2020

# Brain Decoders

Sangil Lee
*University of Pennsylvania*

# Brain Decoders


## Abstract

Functional neuroimaging has opened the door for examining in vivo neural processes for human perception, cognition, and behavior. Naturally, we wonder if we can 'read the mind' from brain images. Typical methods of fMRI are not apt for this, however, as it is aimed at answering 'given a mental state, which brain region is active?' rather than 'given a brain activity, which mental state?'. As a solution, I examine a method using the entire brain to build 'brain decoders' that can empirically measure mental processes. Since different mental processes are unlikely to share the exact same pattern of whole-brain activity, whole-brain decoders could improve specificity to mental processes. Simultaneously, because it recruits multiple regions' signals, whole-brain decoders could also improve our sensitivity to detect mental processes. In the first study, I address the statistical and substantive difficulties of whole-brain decoders and propose a novel algorithm that can overcome them. In the second study, I build a whole-brain decoder of valuation across two economic decision-making tasks and show how a post-hoc analysis of the decoder can yield insight into signal relationships between different regions. In the third study, I showcase how empirical measurements of mental processes can be used in psychological research. Existing theories have posited that people discount delayed rewards because they are imagined less vividly than immediate rewards. I provide neural evidence to this claim by building a whole-brain decoder of imagination vividness on one dataset and show that it can also predict temporal delay in two other datasets of delay discounting task. This dissertation, taken together, shows that whole-brain decoders can be an easy analysis to implement that captures unique neural signatures of tasks and provide measurements of mental processes and constructs.



## Degree Type

Dissertation

## Degree Name

Doctor of Philosophy (PhD)

## Graduate Group

Psychology

## First Advisor

Joseph W. Kable

## Second Advisor

Eric T. Bradlow

## Keywords

Brain reading, fMRI decoder

## Subject Categories

Advertising and Promotion Management | Cognitive Psychology | Marketing | Neuroscience and Neurobiology

BRAIN DECODERS

Sangil Lee

A DISSERTATION

in

Psychology and Marketing

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

_____

Joseph W. Kable
Baird Term Professor
Professor of Psychology

Co-Supervisor of Dissertation

_____

Eric T. Bradlow
K.P. Chao Professor,
Professor of Marketing, Statistics,
Economics, and Education

Graduate Group Chairperson

_____

Sara Jaffee
Professor of Psychology

Graduate Group Chairperson

_____

Nancy Zhang
Ge Li and Ning Zhao Professor
Professor of Statistics

Dissertation Committee

Russell Epstein, Professor of Psychology

Michael L. Platt, James S. Riepe University Professor, Professor of Marketing, Neuroscience, and Marketing

# ABSTRACT

BRAIN DECODERS

Sangil Lee

Joseph W. Kable & Eric T. Bradlow

Functional neuroimaging has opened the door for examining in vivo neural processes for human perception, cognition, and behavior. Naturally, we wonder if we can 'read the mind' from brain images. Typical methods of fMRI are not apt for this, however, as it is aimed at answering 'given a mental state, which brain region is active?' rather than 'given a brain activity, which mental state?'. As a solution, I examine a method using the entire brain to build 'brain decoders' that can empirically measure mental processes. Since different mental processes are unlikely to share the exact same pattern of whole-brain activity, whole-brain decoders could improve specificity to mental processes. Simultaneously, because it recruits multiple regions' signals, whole-brain decoders could also improve our sensitivity to detect mental processes. In the first study, I address the statistical and substantive difficulties of whole-brain decoders and propose a novel algorithm that can overcome them. In the second study, I build a whole-brain decoder of valuation across two economic decision-making tasks and show how a post-hoc analysis of the decoder can yield insight into signal relationships between different regions. In the third study, I showcase how empirical measurements of mental processes can be used in psychological research. Existing theories have posited that people discount delayed rewards because they are imagined less vividly than immediate rewards. I provide neural evidence to this claim by building a whole-brain decoder of imagination vividness on one dataset and show that it can also predict temporal delay in two other datasets of delay discounting task. This dissertation, taken together, shows that whole-brain decoders can be an easy analysis to implement that captures unique neural signatures of tasks and provide measurements of mental processes and constructs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

# CHAPTER 1 - INTRODUCTION

We, humans, have yet to figure out how exactly this hemisphere, barely 6~7 inches in diameter, sitting in a insulated chamber behind our eyes, can solve innumerable complex tasks pivotal to our survival and yet still have some and bandwidth time left over to ponder about its existence. Nevertheless, with recent advancements in neuroimaging, we have made significant *head*way in statistically quantifying the degree of relationship between local brain activities and a menagerie of human perceptions, cognitions, and behaviors. As obvious as it sounds to us today, perceiving objects with our eyes is associated with increased activity in the visual cortex, and movement of our body parts is associated with activity in the motor cortex. Perhaps less obviously, the degree to which we place value on a given object or an option is reflected in the activity of ventromedial prefrontal cortex (Bartra, McGuire, & Kable, 2013; Kable & Glimcher, 2007; Karmarkar, Shiv, & Knutson, 2015; Knutson, Taylor, Kaufman, Peterson, & Glover, 2005; D. J. Levy & Glimcher, 2012; I. Levy, Lazzaro, Rutledge, & Glimcher, 2011; Plassmann, O'Doherty, & Rangel, 2010, 2007), and our language processing is reflected in the temporoparietal junction (DeWitt & Rauschecker, 2012, 2013; Price, 2012). These findings come from many experiments that manipulated the inputs to our brains (or the outputs from our brains), observing a localized activity in a brain region, and ascribing to it, a potential function that is related to the inputs and outputs that were manipulated.

However, we soon learned that it is often difficult to assign a single role to a given region. First, any given stimuli or task will likely result in activity of multiple brain regions. Second, some of those regions, especially those in frontal areas of the brain, will likely appear in other seemingly unrelated tasks as well. The ventromedial prefrontal cortex not only reflects valuation signals, but also reflect movement directions in spatial navigation (Constantinescu, O'Reilly, & Behrens, 2016; Doeller, Barry, & Burgess, 2010); the temporoparietal junction, while

famous for including Wernicke's area for language processing, is also known to be active when we put ourselves in other peoples' shoes to empathize or to simulate others' beliefs (Kanske, Böckler, Trautwein, Lesemann, & Singer, 2016; Kanske, Böckler, Trautwein, & Singer, 2015; R. Saxe & Kanwisher, 2003; Rebecca Saxe & Powell, 2006; Young, Cushman, Hauser, & Saxe, 2007; Young, Dodell-Feder, & Saxe, 2010). The most notorious example comes from a non-peer-reviewed study, published in the New York Times, that claimed that people literally love their iPhones, because seeing their iPhones results in activity in the insula, which has been associated with feelings of love before. Around 50 neuroscientists wrote a disappointed letter to the editor that read "The region that he points to as being "associated with feelings of love and compassion" (the insular cortex) is a brain region that is active in as many as one third of all brain imaging studies." Known as the reverse inference problem, inferring a mental state given brain activity is more complex for regions that are active in a variety of tasks.

It is here that this dissertation attempts to venture a path forward, with a simple idea: instead of inferring mental states from one region, we should try to infer mental states using the entire brain and the pattern of activity within. The activity in the ventromedial prefrontal cortex may be caused by many things, but if it's co-activated with ventral striatum and posterior cingulate cortex, there's a good chance that it is a valuation signal. Temporoparietal junction may be involved in multiple activities, but if it's joined by middle and inferior temporal gyri and supplemental motor area, that is likely someone producing speech. By combining multiple regions' signals across the brain, we can reduce the chances of making a wrong inference (i.e., increase the specificity of inference), while simultaneously increasing our power to detect a given mental state even when one region is missing (i.e., increase the sensitivity of inference).

It's important to acknowledge, before proceeding, that the idea of whole-brain prediction is not new. It's just unpopular. For example, Wager et al. (2013) has combined principal component analysis with penalized regression to build a whole-brain decoder of pain. Smith and colleagues used whole-brain data in a penalized regression model to decode peoples' preferences between items (Smith, Douglas Bernheim, Camerer, & Rangel, 2014). Kragel and Labar used partial least squares discriminant analysis to develop categorical predictors of 7 emotional states (Kragel & LaBar, 2014). Grosenick, Greer and Knutson developed a spatially penalized version of the elastic net regression to build predictors of consumer purchases (Grosenick, Greer, & Knutson, 2008). Despite these uses, two reasons make such whole-brain approach less favored than partial-brain approaches. A first difficulty is interpretability. Whole-brain predictors, as it is currently implemented, are often difficult to interpret. For neuroscientists, who want to know the brain regions that are predictive of certain behavior or mental state, it does not suffice to know that the prediction is possible via some black-box algorithm; rather the resulting predictor must shed light on identifying key regions and setting them apart from regions that are not predictive. Unfortunately, in the absence of prediction algorithms specialized for fMRI, off-the-shelf methods for dealing with high-dimensional data are often not suited for identifying clusters of signals that can be localized as a region. A second difficulty is that dealing with all brain regions simultaneously is statistically and computationally challenging. With many variables but few observations, researchers need to rely on modern methods such as machine learning to produce predictors; unfortunately, many such methods take considerable amount of computational resources which render the method less accessible than partial-brain prediction approaches.

Naturally, the first step of this dissertation is to provide a novel method, suited for fMRI, that can build interpretable whole-brain decoders in a short time with little computational resources. In the first study, I introduce a new algorithm, named Thresholded Partial Least

Squares (TPLS), that exploits analytical properties of partial least squares algorithms to build high-dimensional predictors with very little computation. I demonstrate, in a large fMRI dataset of intertemporal choice and risky choice, that TPLS is much faster to train compared to other methods while still having high predictive power. Most importantly, I compare the resulting whole-brain decoders from TPLS and other extant methods, using both simulation and real data, to show that TPLS whole-brain decoders lead to clustered localized signal that identifies key predictor regions.

The second study highlights what we can learn from the whole-brain predictor created from study 1 by analyzing it in detail. When examined, I found that the whole-brain predictor of value contains many regions with negative regression coefficients, despite the fact that no brain region has signal that is negatively correlated with value. From this observation, we get the clue that there may be common noise in the value signal that hinders prediction of choice; consequently, the whole-brain predictor pits some of the regions against each other to cancel out this common noise while retaining the value signal. Based on this empirical finding, I found that while peoples' choices were not affected by the values of options in past trials, the current neural value signal was affected by them. This history dependency, which is only found in neural signal of value but not in value-based behavior, becomes the 'noise' that the whole-brain predictor tries to filter out. This study demonstrates, above the important finding that neural correlates of value have history dependency, that we can learn more about the nature of the neural signals by examining the whole-brain predictor.

The third and final study of this dissertation showcases how 'decoding the brain' can be put to use in psychological research. There are several existing theories such as construal level theory arguing that the reason people discount delayed rewards is because delayed rewards, due to their temporal distance, are imagined less vividly (or are construed more abstractly) than

immediate rewards (Liberman & Trope, 2014; Rick & Loewenstein, 2008; Trope & Liberman, 2010). The vividness of imagination is, needless to say, a very subjective experience that can be difficult to probe with behavioral measures such as self-reports. Furthermore, asking participants about how vividly they considered the delayed reward can easily lead to experimenter expectancy bias. This is where a 'neural read' of the brain can be immensely useful. I provide neural evidence for these theories by first constructing a brain-decoder of 'imagination vividness' on a dataset where participants were instructed to imagine about vivid and non-vivid scenarios. Then, on two completely separate delay discounting datasets, I show that the decoder's reading of peoples' brain when performing delay discounting task, is negatively correlated with delay of the reward, suggesting that rewards that are farther in time are indeed imagined less vividly.

The three studies in this dissertation makes a strong argument to researchers to consider using whole-brain decoders in their research. The first study provides a fast, easy to use algorithm with interpretable results in two statistical languages (MATLAB & R). The second study shows how examining the whole-brain decoder can lead to further understanding of the neural signals involved. The third study shows that whole-brain decoders can measure mental processes and constructs that have been traditionally difficult to assess.

CHAPTER 2 – Building a Brain Decoder: Application of Thresholded Partial Least Squares

Sangil Lee, Eric T. Bradlow, and Joseph W. Kable

**Abstract**

Recent neuroimaging research has shown that it's possible to decode mental states and predict future consumer behavior from brain activity data (a time-series of images). However, the unique characteristics (and high dimensionality) of neuroimaging data, coupled with a need for neuro-scientifically interpretable models, has largely discouraged the use of off-the-shelf prediction methods. Instead, most neuroscientific research uses only "regionalized" (partial-brain) data, leading to a loss of potential information, and simple methods to reduce the computational burden and to improve interpretability (i.e., localizability of signal). Here we propose a novel approach that can build whole-brain neural decoders (using the entire data set and capitalizing on the full correlational structure) that are both interpretable and computationally efficient. We exploit analytical properties of partial least squares algorithm to build a regularized regression model with variable selection that boasts (in contrast to most statistical methods) a unique 'fit-once-tune-later' approach where users need to fit the model only once and can choose the best tuning parameters post-hoc. We demonstrate its efficacy in a large neuroimaging dataset against off-the-shelf prediction methods and show that our new method scales exceptionally with increasing data size, yields more interpretable results, and uses less computational memory, while retaining high predictive power.

## Introduction

Functional neuroimaging has allowed researchers to empirically examine the relationship between brain activity and various mental processes. In particular, functional magnetic resonance imaging (fMRI) has shown great success in correlating brain activity with various behaviors and cognitive processes. Brain activities are measured in small grids, in 3D units known as voxels (i.e., volumetric pixels), and a typical fMRI image of the brain can have anywhere between 40,000~200,000 voxels. Therefore, while a single scan of brain activity contains a large number of variables (i.e. a classic large P problem), the number of subjects (observations) are relatively much smaller (i.e. a classic small N problem) as it needs to be collected via an in-person experimental task, conducted inside an MRI machine, usually operated by a trained technician. Therefore, the regime of predicting cognitive states and behavior using voxels in fMRI is in the domain of $P \gg N$, where the number of predictors far outweigh the number of observations. This has led most studies in neural prediction to focus on small cutouts of the brain (lowering P), wherein researchers have a priori interest/scientific theory.

But the sheer number of predictors, by itself, is not the only reason why most neural prediction studies focus on partial-brain prediction. For neuroscientific research, it's less important that prediction of behavior is possible than to know which brain area makes it so. Modern methods for high-dimensional data are not too helpful for the researcher unless the predictor is interpretable and can lead to scientific insights about the brain. Concordantly, instead of whole-brain prediction, a common method has been to use 'searchlight analysis' that iterates through the brain, taking a handful of local voxels at a time, to identify which regions can predict the given behavior or cognitive states above chance (Etzel, Zacks, & Braver, 2013).

However, ideally, neural prediction should be done using *all* the voxels in the brain for several reasons. First, a whole-brain predictor can provide higher sensitivity than ROI methods to decode mental states by combining signals across the entire brain. Second, it can also provide higher specificity by providing a unique predictor for each mental state, while any given region of the brain may co-activate with many different mental states. Third, a whole-brain predictor can give insight into how different regions' signals are combined, much like how regression can give different answers than pairwise correlations. These benefits, coupled with recent advent of large-scale datasets and collaborations in neuroimaging, has opened doors for construction of generalized whole-brain decoders. For example, Wager et al. (2013) has constructed a whole-brain signature of pain that predicts the degree of pain an individual feels. Smith et al. (2014) and Lee, Lerman, & Kable, (2019) has constructed whole-brain predictors of valuation. In particular, Lee et al. (2019) shows how a whole-brain predictor can reveal complementary relationships between different brain regions that cannot be assessed from region-based methods.

Unfortunately, an interpretable whole-brain decoder is not readily achievable with off-the shelf statistical methods. As expected, a first difficulty is interpretability: a good whole-brain predictor should distinguish *regions* that are predictive versus those that are not. As a counterexample, predictors built from LASSO (a penalized regression approach) results in predictive maps with scattered sparkles of coefficients across the brain rather than any interpretable clusters or regions (Grosenick, Klingenberg, Katovich, Knutson, & Taylor, 2013). This is due to the feature of penalized regression methods to select the few most predictive variables instead of clusters of correlated variables. On the other hand, using PCA in conjunction with penalized regression models (PCR-LASSO), Wager et al., (2013) was able to provide whole-brain prediction maps that had clustered coefficients for voxels that delineates which regions are positively predictive and which regions are negatively predictive. This natural clustering of

coefficients is possible due to the fact that the extracted principle components are correlated with neighboring voxels to a similar degree. However, PCR-LASSO does not provide any voxel selection measure and hence results in a prediction map where every voxel is given a coefficient (in contrast to a thresholded approach as described in this research). In sum, previous approaches to whole-brain methods have had difficulty in providing 1) regionally clustered coefficients, and 2) selectivity of important vs. unimportant regions.

A second difficulty is computational efficiency, especially with regards to scaling with larger datasets. As mentioned, since neuroimaging data has a substantial number of predictors, purely likelihood-based approaches often face the problem of calculating gradients for a large number of variables. Adding to the burden, modern models need to be fit multiple if not hundreds of times to find the best tuning parameter (e.g., tuning parameter $\lambda$ in LASSO that controls variable selectivity). These problems are computationally challenging even in average sized fMRI datasets, which is why previous whole-brain predictors used down-sampled images (i.e. coarser) with fewer voxels for prediction (e.g., Grosenick et al., 2013; Wager et al., 2013). In contrast to purely likelihood-based methods, data-reduction approaches such as PCA can help in immensely reducing the number of variables and thereby reducing the model fitting time. However, in larger datasets, PCA itself can become a bottleneck for computation time and memory usage as it requires computation of the variance-covariance matrix of predictors. These computational costs prevent the widespread use of whole-brain methods in neuroimaging, especially for those without access to computational clusters.

Here we propose a novel method, thresholded partial least squares (TPLS), that provide interpretable whole-brain predictors that are computationally efficient enough to run on personal computers for most datasets. TPLS exploits analytical properties of a modified partial least squares algorithm to offer a unique ‘fit-once-tune-later’ approach where the user fits the model

only once and then evaluates the best tuning parameter as many times as needed without re-fitting the model. This is in stark contrast to most, if not all, modern methods that require re-fitting the model for every tuning parameter. Here, we describe the algorithm and showcase its performance against other methods in a large neuroimaging dataset. Furthermore, we provide the TPLS package online for MATLAB at (https://github.com/sangillee/TPLSm) and for R at (https://CRAN.R-project.org/package=TPLSr) thus, we hope, a practical tool for others.

## Methods

### *Overview*

TPLS, like many modern regression methods, requires two (training) estimation steps: fitting and parameter tuning. In the fitting step, the end goal is to calculate the coefficient and the z-statistic of each original variable (e.g., voxel) – similar to the coefficients and t-statistics one would get for each variable in multiple regression (**Fig. 1-1**). In detail, the fitting step first extracts the PLS components that maximally explain the covariance between X and Y and then regresses Y against the components[1]. The resulting coefficients and the t-statistics of the components are then back-projected into the original variable space so that each voxel would have a coefficient and a z-statistic. In the tuning step, the user decides on how many PLS components to keep, and how many voxels to keep based on the (researcher chosen) z-statistic threshold. For example, using cross-validation, the user may find that a TPLS model that uses the first 6 PLS components and retains 50% of the original variables provides the highest out-of-sample cross-validation performance. Intuitively, the number of PLS components controls the

[1] Technically, the calculation of the components and their regression coefficients are done simultaneously based on analytical methods without having to perform regression.

degree of data reduction while the thresholding level controls variable selection, and, as desired here, improves interpretability.

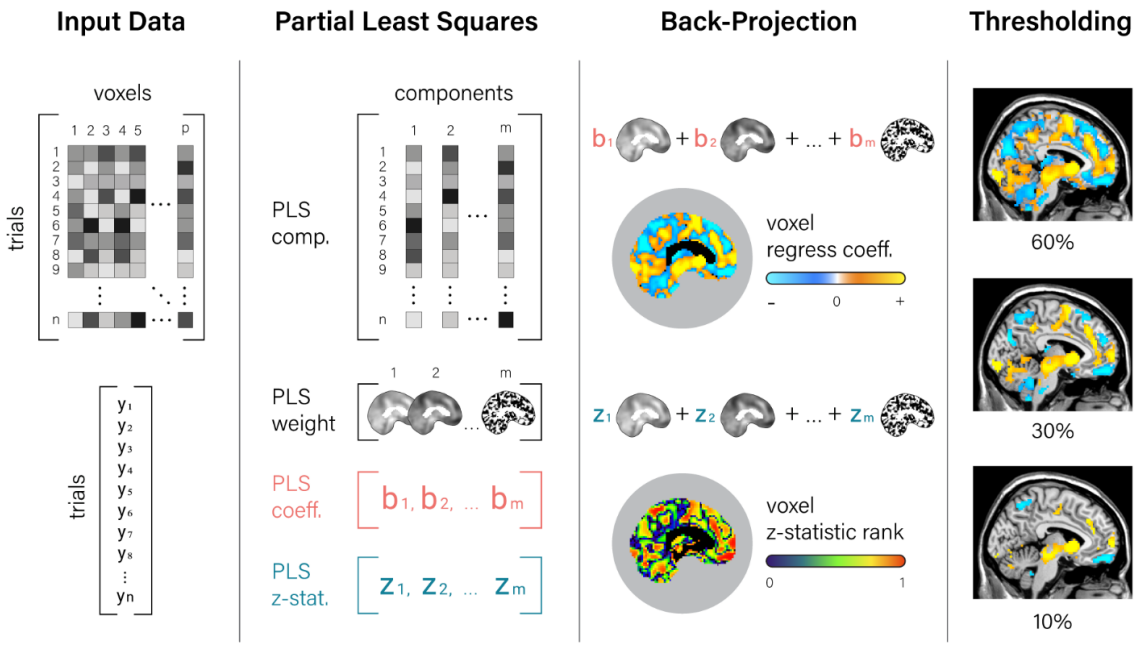


**Figure 1 - 1. Schematic of TPLS fitting procedures.**

TPLS model fitting first requires extracting the partial least squares components from the predictor matrix and obtaining the components' back projection maps (PLS weight), their regression coefficients, and their z-statistic. Next, the regression coefficients and the z-statistics are then back-projected into the voxel space using the weight maps, thereby yielding a whole-brain coefficient map and a whole-brain z-statistic map, which is then ranked in absolute size to determine voxel importance. Finally, the coefficient map is thresholded based on the z-statistic rank map to select voxels that are the most important.

The key computational benefit of TPLS comes from the 'fit-once-tune-later' feature. The user can choose among infinitely many tuning parameter combinations without having to re-fit the model, as all the information required is already calculated in a one-time fitting. This is because once a TPLS model with *m* components has been fit, all the models with fewer components are also available (i.e., a 1-component, 2-component, …, *m*-component model). Since PLS components are all orthogonal to each other, their regression coefficients do not

change based on the components you decide to keep, thereby allowing the user to choose the necessary components without re-fitting.

TPLS falls into the data-reduction category of models, just like PCR-LASSO. However, PLS provides two key benefits over PCA. First, PLS computation only requires vector multiplications, which are fast and memory efficient, while PCA requires matrix singular value decompositions, which grow quadratically with data. Second, PLS components are ordered in terms of the explained covariance of X on Y, while PCA components are ordered only in terms of variance explained in X. A standing criticism of using PCA for data reduction is that while PCA components are great for describing X, they are not necessarily relevant in predicting Y (Lever, Krzywinski, & Altman, 2017).

*TPLS Algorithm - fitting*

The fitting algorithm for TPLS is a combination of three parts: a modified SIMPLS algorithm for PLS (de Jong, 1993), back-projection, and calculation of z-statistics. **Fig. 1-2** shows the algorithm of TPLS. The one modification that we make to the SIMPLS algorithm is simply in normalizing the PLS components to have weighted unit variance (step 4 in **Fig. 1-2**). This facilitates computation of z-statistics later in the algorithm (step 7 in **Fig. 1-2**). Below we detail the back-projection and the z-statistic calculation of TPLS as the rest are typical procedures of a SIMPLS algorithm.

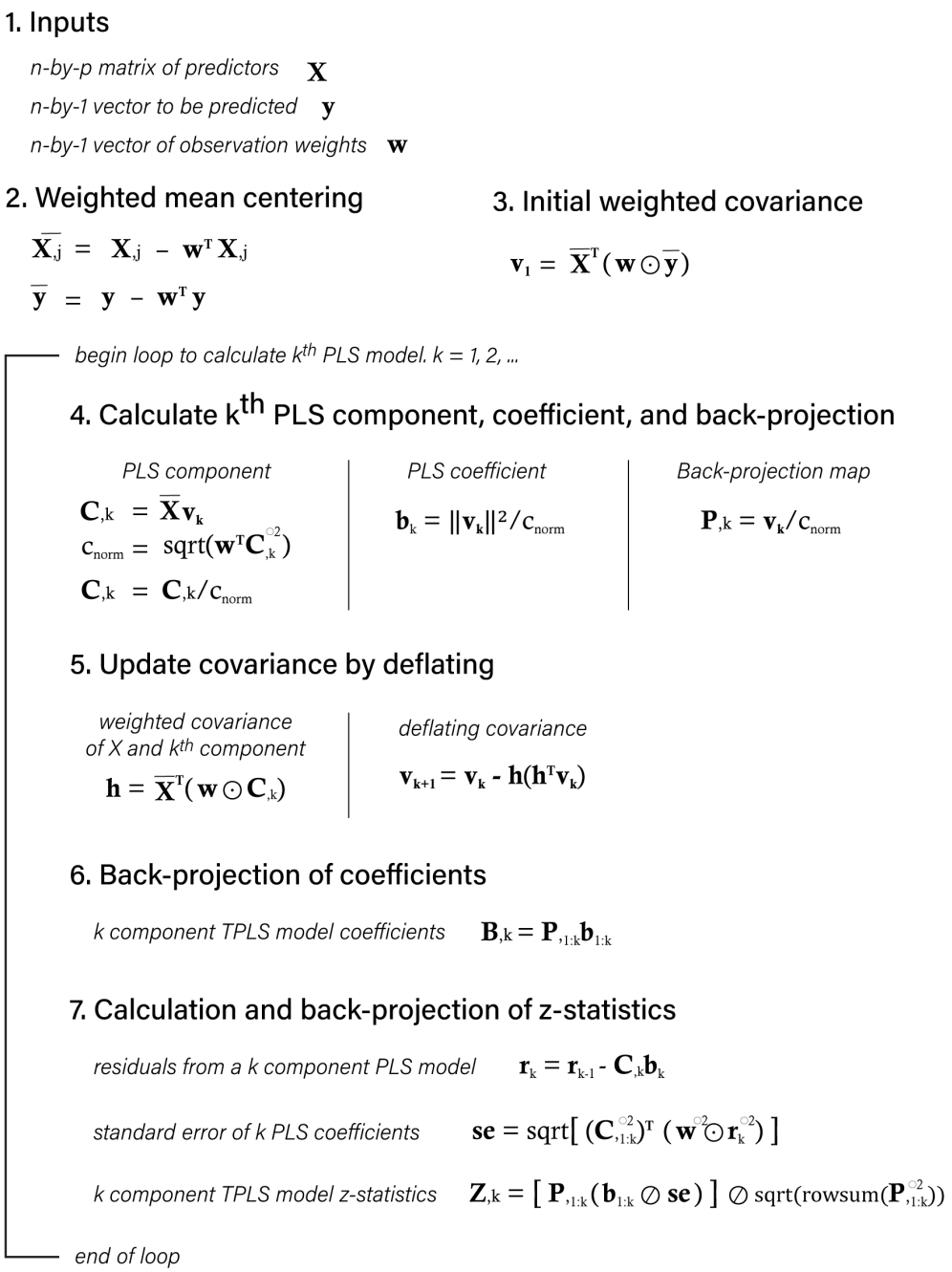


**Figure 1 - 2. Summary of TPLS fitting algorithm.**

Matrices are denoted with bold capital letters, vectors with bold lowercase letters, and scalars with non-bolded lowercase letters. $\odot$ denotes Hadamard product (elementwise multiplication), $\oslash$ denotes elementwise division, and $\circ 2$ in the exponent denotes elementwise squaring.

*Back-projection* (step 6 in **Fig. 1-2**). After PLS components have been calculated (up to the $k^{th}$ component), we now have a *k*-component PLS regression model. To improve the interpretability of this PLS model, we can convert the PLS regression coefficients into the original voxels' coefficients. Since PLS components are created via weighted sums of original variables (i.e., component = weight * variables), one can simply multiply the PLS coefficient by the weights first to create back-projected coefficients (i.e., coefficient * component = coefficient * weight * variables = back-projected coefficient * variables). This expresses the PLS regression in terms of each voxel's coefficients, which can make the predictor easier to interpret by identifying which regions are positively or negatively predictive of behavior and mental states. This back-projection is also used in the PCR-LASSO method used by Wager et al. (2013), but with PCA components rather than PLS.

*Z-statistic calculation* (step 7 in **Fig. 1-2**). We then calculate the z-statistic of each voxel as a measure of variable importance. We start by calculating the heteroscedasticity-consistent standard errors (also known as sandwich estimators; White, 1980):

$$\mathbf{Var(b)} = (\mathbf{C^T diag(w) C})^{-1}\mathbf{C^T diag(w) M diag(w)^T C}(\mathbf{C^T diag(w) C})^{-1}. \quad (1)$$

where **w** denotes the observation weights, and **M** denotes the variance-covariance matrix for the observations. Here is where our modification to the SIMPLS algorithm come in handy. Since the PLS components (matrix **C)** are all orthonormal (in weighted space), $\mathbf{C^T diag(w) C}$ becomes an identity matrix, which cancels out the 'breads' of the sandwich and leaves us with $\mathbf{Var(b)} = \mathbf{C^T diag(w) M diag(w)^T C}$. Since we only need the diagonals of the variance-covariance matrix, we can express the standard error estimates concisely as the following:

$$\mathbf{se(b)} = \sqrt{\left((\mathbf{C}^{\circ 2})^{\mathbf{T}}(\mathbf{w}^{\circ 2} \odot \mathbf{r}^{\circ 2})\right)}. \quad (2)$$

where $\mathbf{r}^{\circ 2}$ denotes the squared residual vector. The t-statistics (which is close to a z-statistic with sufficient observations) can be then calculated by simple element-wise division of $b$ by $\mathbf{se(b)}$. Let this vector be denoted **z**. Then, we back-project the z statistic like the coefficients, and then normalize them so that they all have unit variance:

$$\left[\mathbf{P}_{,1:k}\left(\frac{\mathbf{b}_{1:k}}{\mathbf{se}}\right)\right] / \sqrt{\mathbf{rowsum}\left(\mathbf{P}^{\circ 2}_{,1:k}\right)} \quad (3)$$

where **rowsum** denotes the vector that is the row sum of a matrix. This summarizes the fitting procedure of TPLS.

*TPLS Algorithm - tuning*

After the model fitting is complete, one can retroactively choose the two tuning parameters, number of components and the thresholding level, based on cross-validation performance. **Fig. 1-3** shows an example cross-validation performance surface as a function of the number of PLS components and the thresholding level. Because the model does not have to be re-fit, researchers can examine the predictor at various thresholding levels and assess the tradeoff between predictive performance.

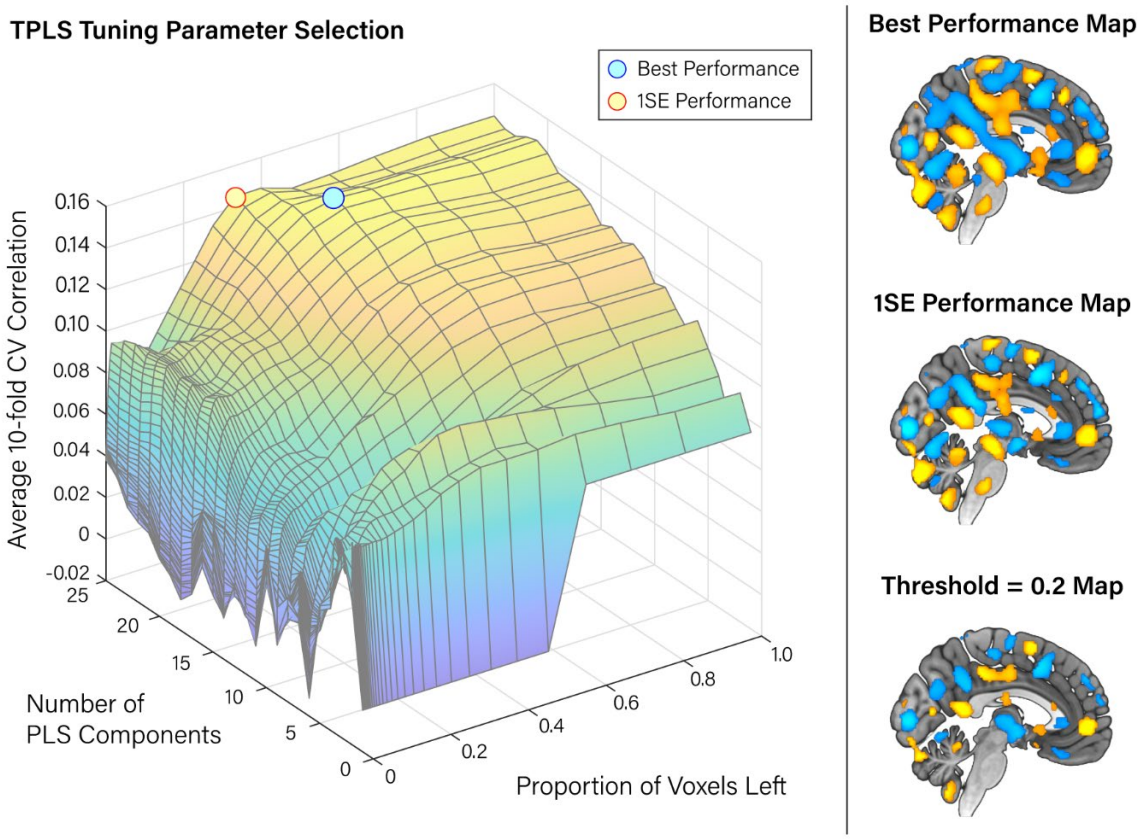


**Figure 1 - 3. Example TPLS model tuning.**

The left panel shows an example cross-validation performance surface as a function of the two tuning parameters of TPLS: number of PLS components (1~25) and proportion of voxels left (0~1). The highest CV performance point is marked with a blue dot with the corresponding whole-brain predictor shown on the right top panel. Additionally, a model with fewer voxels but within 1 standard error of the best model's performance is indicated with a yellow dot with the corresponding map shown on the right middle panel. The right bottom panel is shown to illustrate how the number of remaining voxels with coefficients reduce as the proportion of voxels left are reduced.

After the tuning is complete, there is one more step that may be useful in some scenarios: post-fitting of bias (intercept). Since some variables are removed during the thresholding stage, the intercept should be re-fitted after thresholding. Let's say that we chose to evaluate a model with *j* components, thresholded at 70% (removing 70% of variables). Then, the coefficients are $\mathbf{B}_{\cdot,j}$ multiplied by index vector $\mathbf{d}$ where $\mathbf{d}_i = 1$ if the voxel's rank is in the top 30% and 0

otherwise. Then the new intercept is simply the difference between the weighted means of **X** and **y**:

$$b_0 = \mathbf{w}^T\mathbf{y} - \mathbf{w}^T\mathbf{X}(\mathbf{B}_{.j} \circ \mathbf{d}). \quad (4)$$

*Neuroimaging Dataset*

We used a large neuroimaging dataset from Kable et al., (2017) to empirically compare TPLS against other extant whole-brain methods (PCR-LASSO and LASSO), and ROI-based predictions. We chose this dataset as it was the most readily available large-scale dataset with whole-brain coverage that can be used to decode behavior from brain activity levels. Participants completed two experimental decision-making tasks: one an intertemporal choice and one a risky choice, both very common in the domain of social science (psychology, economics research, e.g., Green & Myerson, 2004; Kahneman & Tversky, 1979; Samuelson, 1937) and their interaction with neuroscience (e.g., Jung, Lee, Lerman, & Kable, 2018; Kable & Glimcher, 2007). In intertemporal choice, participants made choices between a smaller immediate monetary amount of $20 and a larger but delayed monetary amount (e.g., $40 in 30 days). This choice paradigm allows one to assess the "time value of money", hence its name "intertemporal choice". In risky choice, subjects made choices between a smaller certain monetary amount of $20 and a larger but probabilistic monetary amount (e.g., $40 with 60% probability of winning). In both tasks the larger amount varied from trial to trial as well as the delay and the risk in each respective task, while the smaller monetary option was always fixed at $20. Only the larger monetary option was on the screen while the smaller $20 was not; participants made accept/reject choices based on whether they'd prefer the larger monetary option on the screen or the smaller monetary option. Because the value of one of the options were always constant, Kable et al. (2017) was able to find

valuation signals in the brain that correlated with the utility of the varying option that was shown on the screen. Based on this result, we sought to create a whole-brain predictor of choice that can use these valuation signals to predict whether the participant will accept the option on the screen or reject it. Details about removed participants, fMRI image acquisition protocols and preprocessing details are provided in the **supplemental materials**. In total, the dataset gave us a total of 61,038 trials (observations) and 184,319 voxels (variables) across 531 task sessions (264 intertemporal choice sessions, 267 risky choice sessions), which we'll treat as 531 participants in this paper, as our goal is not in making substantive, or comparative, conclusions about the tasks.

*Computation comparisons*

We first assess the scalability of each whole-brain prediction method – LASSO, PCR-LASSO, and TPLS – by comparing their model fitting time and RAM usage at varying training dataset sizes (8, 16, 32, 64, 128, 256, and 512 participants). In each dataset size (e.g., 8 subjects), half of the data was drawn randomly from the risky choice dataset (i.e., 4 subjects) and the other was drawn randomly from intertemporal choice dataset. Each model was fitted using 10-fold cross-validation (CV). The training data was divided into 10 equal sized blocks and the model was fitted on 9 of the blocks and tested on the left-out block. This was repeated 10 times to assess cross-validation performance. Then, the tuning parameter that yielded the highest CV performance was chosen and used to train the final predictor using all training data.

For LASSO, we used GLMNET for MATLAB (Friedman, Hastie, & Tibshirani, 2010; Qian, Hastie, Friedman, Tibshirani, & Simon, 2013), which is arguably the fastest package for

fitting LASSO thanks to its use of regularized path and FORTRAN coding[2]. We used default tuning parameter search, which used 100 lambda values. For PCR-LASSO, we stick with the original approach in the Wager et al., (2013) paper by extracting 200 components from all data, using 10-fold LASSO logistic regression to find the useful components, and subsequently running an unpenalized logistic regression using only the selected components. For TPLS, in each of the 10 folds, we extracted 25 PLS components and built the TPLS model. Then, during cross-validation we chose the best number of PLS components and threshold levels. Each whole-brain method was fitted 400 times at each dataset size, each time randomly selecting the training data. All computations were performed on a large-scale computation cluster at the University of Pennsylvania (https://www.med.upenn.edu/cbica/cubic).

*Predictive power comparisons*

For predictive performance, we compared the out-of-sample predictive performances of the predictors built above. After the prediction model was fitted using 10-fold cross validation in the training data (e.g., 32 subjects), all unused data (e.g., 531-32=499 subjects) served as out-of-sample testing dataset. Per-subject correlation and area under the ROC curve (AUROC) were averaged across the out-of-sample participants to get an average estimate of out-of-sample prediction performance. The AUROC of receiver operating characteristic can be understood as 'threshold-free accuracy', ranging from 50% to 100%, as it has been shown to equal the probability that the model will accurately give a higher score to a randomly chosen positive case than a randomly chosen negative case (Fawcett, 2006). We also added two commonly used ROI-

[2] The original GLMNET package for MATLAB could not import a dataset size of as large a magnitude as in this study because the FORTRAN API with MATLAB was written in 32-bit architecture; we have updated the FORTRAN code ourselves to 64-bit architecture to circumvent this issue; the updated package is provided here: https://github.com/sangillee/GLMNET64MATLAB

based prediction methods to the comparison of predictive power: ROI-average, and ROI-multivariate. ROI-average is simply taking the average of all voxel activities within a designated ROI to make predictions; concordantly, ROI-average does not require fitting a model. ROI-multivariate, on the other hand, uses the voxels in the ROI to build a predictor. While several methods can be used, here we used LASSO to keep make comparisons with our whole-brain methods easier. We used ROIs from a meta-analysis by Bartra, McGuire, & Kable (2013), which examined around 150 neuroimaging studies on valuation signals and identified two ROIs that consistently showed correlated activity with valuation: ventral striatum and ventromedial prefrontal cortex.

*Interpretability comparisons*

We compared the interpretability of three whole-brain methods (LASSO, PCR-LASSO, TPLS). Using the same fitting procedures as before (10-fold cross-validation), TPLS and PCR-LASSO was fit using the entirety of the data (531 participants). LASSO, however, was computationally too slow to fit using the entire dataset and was only fit with a subsampled 64 participant dataset. We visually compared the resulting whole-brain predictors and the associated areas of the brain to assess the (face/scientific) validity of the identified brain regions.

Additionally, we used simulated data to shed further insight into differences in interpretability of TPLS and extant neural prediction methods. For simulation, we simulated a brain activity signal of a 17x17 voxel grid (total of 289 voxels), of which only a 5x5 grid in the center (25 voxels) carried signal that was predictive of Y, while all other voxels were completely orthogonal to Y (i.e., noise). We achieved this by first randomly generating 290 variables (289 voxels + 1 Y) each with 300 observations from a standard normal distribution. Then, we applied

symmetric orthogonalization such that all 290 columns are orthogonal to each other. The first column of the new matrix was chosen as the predicted variable **Y**, while the other 289 variables became simulations of fMRI noise. To create 25 voxels of predictive voxel signal, we mixed **Y** with 25 of the simulated fMRI noise variables to create 25 signals that were all exactly correlated with Y at r = 0.1. Each column was then z-scored to have unit variance. Finally, we placed the 5x5 signal grid in the center of a 17x17 grid and applied a small amount of 2D Gaussian smoothing (sd = 1 voxel) to simulate the inherent smoothness of fMRI signals. In sum, the resulting dataset was 300 observations of 17x17 voxel grid predictors with only the center 5x5 grid being predictive of Y. This simulated dataset was fit by OLS, LASSO, PCR-LASSO (10 components), and TPLS model (10 components) to compare the resulting pattern of coefficients.

## Results

### *Computation Time*

TPLS shows exceptionally fast model-fitting time compared to the other algorithms especially as the dataset size grows larger (**Fig. 1-4A**). In the largest training dataset size of 512 people (256 sessions of ITC and 256 sessions of risky choice), TPLS took 2 hours and 10 minutes on average to finish 10-fold cross validation training while PCR-LASSO took 2.3 days, which is 25 times slower than TPLS. LASSO was already taking close to 2 days for 64 participants and was too expensive to compute for larger dataset sizes. More importantly however, the differences in computation speed grew larger and larger as dataset size increased. When computation time was divided by the number of participants, we found that while TPLS takes about the same amount of time per-subject (~15 seconds), PCR-LASSO and LASSO shows increasing fitting time per-subject as dataset size increased (**Fig. 1-4B**). This divergence is likely because PCA

requires inversion operations on the variance covariance matrix of X, which will quadratically increase in size until the number of observations match the number of variables. For LASSO, this increase in fitting time is also likely due to calculation of gradients based on matrix operations. TPLS, on the other hand, only require vector calculations, for which only the number of variables is the dominating factor. The speed of TPLS will prove useful for many neuroimaging studies that may examine multiple different behaviors or mental constructs. In addition, since TPLS only has to be fit once (as mentioned), this provides even greater computational benefits.

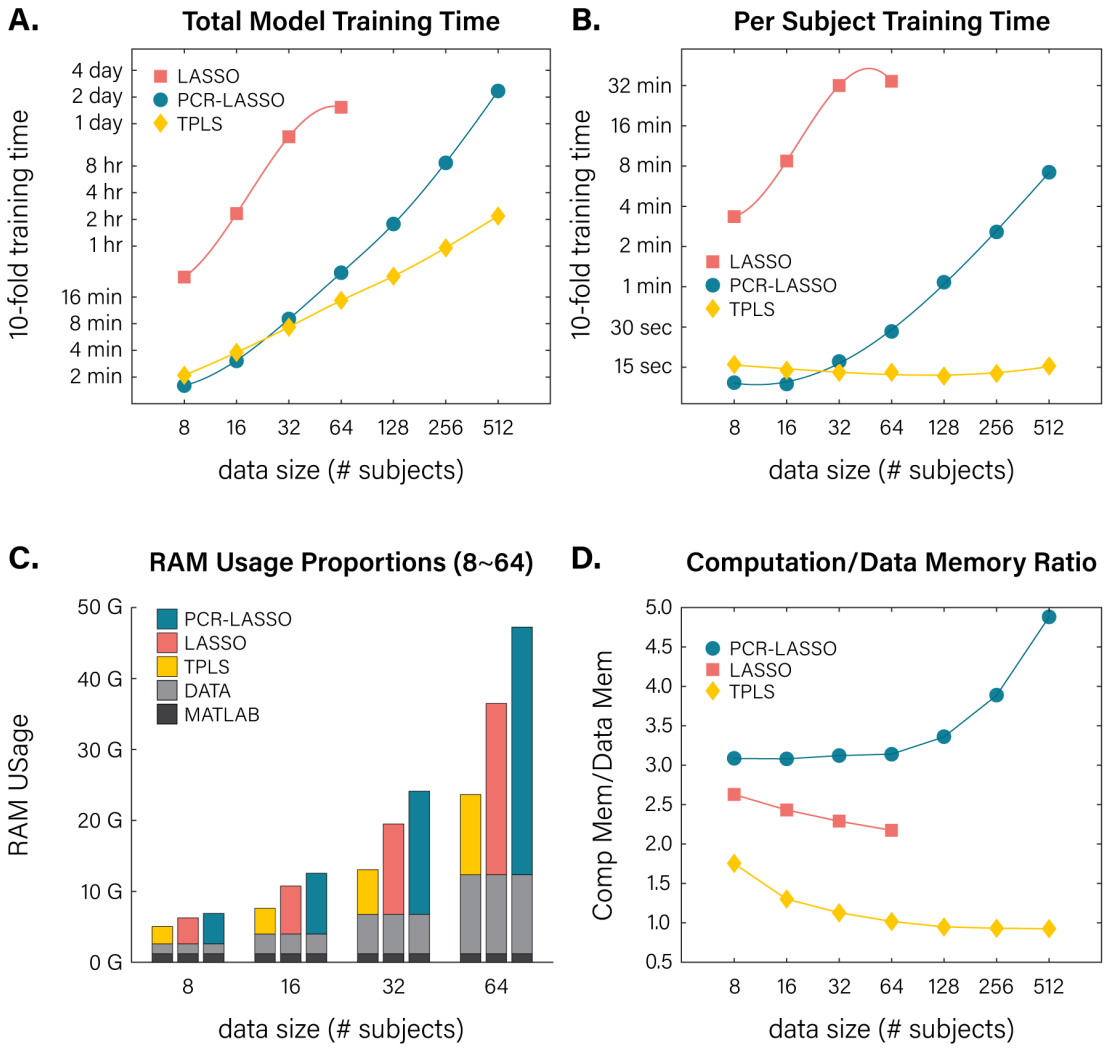


**Figure 1 - 4. Computation resource comparison.**

Panels A and B show 10-fold model training times for each algorithm at various dataset sizes (total time and total time dived by number of subjects, respectively). The lines show best fitting cubic polynomial spline fit. The bottom four panels show RAM usage of each algorithm at various dataset sizes. Panel C shows the memory decomposition of each algorithm (memory for turning on MATLAB, for loaded data, and for computation). Panel D shows the ratio between RAM usage for loading data and RAM usage for computation (colored vs. light grey bar in panel C).

*Memory Usage*

TPLS also uses a very minimal amount of memory compared to other algorithms. (**Fig. 1-4C**). We broke down the memory usage into three parts: default RAM for loading the statistical

program, RAM for loading the data, and RAM for computing the model from the data. The first two is the same across all algorithms as they all need to load the program and the data. The key differences come from the differences in RAM usage for model computation. TPLS used the least amount of computation memory, followed by LASSO and PCR-LASSO.

We also found that TPLS's memory usage was also the most scalable out of all three algorithms (**Fig. 1-4D**). TPLS's computation RAM usage converges to about the same as RAM needed for loading the data (1.75 times -> 0.95 times as dataset size increases). This is likely because TPLS requires a mean-centered copy of the data matrix. Should researchers want to, they can mean-center the data beforehand and use even less memory for TPLS (this feature is available in the provided statistical packages). On the other hand, LASSO's RAM usage for computation ranged from 2.6 times to 2.2 times of data RAM size, while PCR-LASSO's RAM usage was particularly bad as the required RAM for computation increased rapidly as dataset size increased (3 times -> 4.9 times).

*Prediction Performance*

TPLS showed the highest or the second-highest predictive performance across all dataset sizes (**Fig. 1-5**). The ROI-average based prediction method, provided the worst prediction performances across the board, followed by ROI-multivariate method, which built a predictor based on only the voxels within a pre-defined ROI. All three whole-brain predictors provided considerably higher predictive performances that increased with dataset sizes. Of the three methods we compared, PCR-LASSO provided the worst predictive performances on average, except for in the smallest dataset size where LASSO was the worst predictive performance. TPLS showed higher predictive performances compared to PCR-LASSO across all dataset sizes and

compared to LASSO in small dataset sizes (8 ~16). LASSO showed the highest predictive performance in middle dataset sizes (32~64), which would have likely continued to be the highest in larger dataset sizes (128~) had it not been for the lengthy computation time. The differences in predictive power between the whole-brain methods were not as big, however, as the difference between whole-brain methods and ROI-based methods. This illuminates how a whole brain predictor can harness more signals across the brain to provide more power.

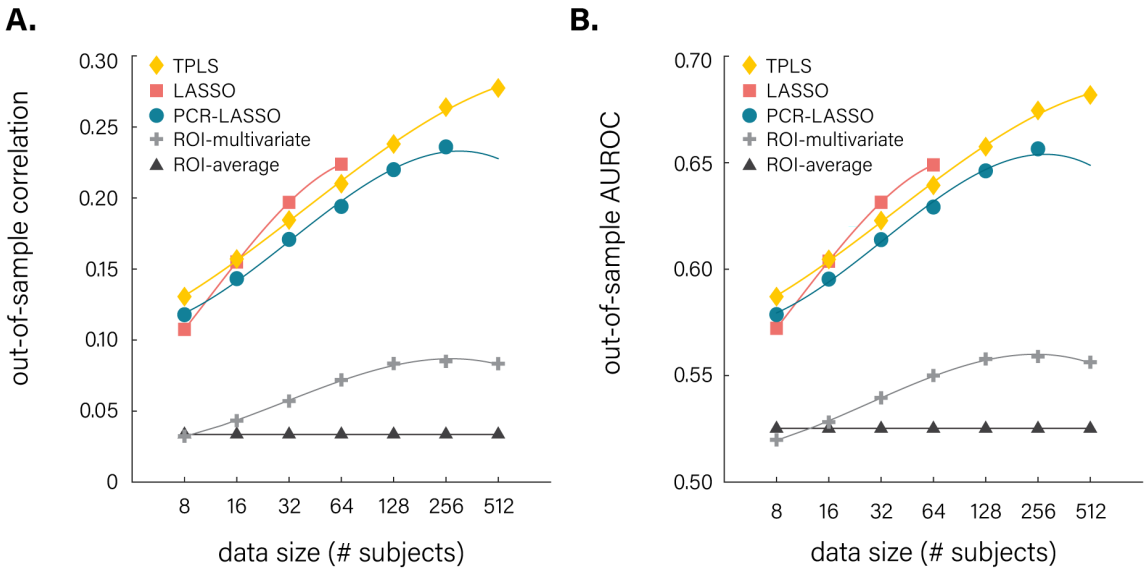


**Figure 1 - 5. Out-of-sample prediction of various algorithms.**

Shown above are out-of-sample prediction performances of five algorithms measured via Pearson correlation (left) and AUROC (right) for predicting value-based accept/reject choices. The lines show best fitting cubic polynomial spline fit.

*Predictor Interpretability*

The resulting whole-brain predictors from various approaches show that different methods lead to different predictors and that TPLS provides predictor maps that are easily

interpretable (**Fig. 1-6**). TPLS results in whole-brain predictors with regionally clustered coefficients and voxel selection (**Fig. 1-6A**). This allows researchers to identify key brain regions of the predictor. On the other hand, PCR-LASSO approach leads to regionally clustered coefficients that is easy to identify but has no voxel selection included (**Fig. 1-6B**). Different from this approach, LASSO predictors were able to select the important predictive voxels from the non-predictive ones; however, the resulting coefficients are scattered such that it is difficult to pinpoint the regions from which the signals originate (**Fig. 1-6C**). A single voxel can be very difficult to interpret as their positions are not always clearly belonging to a region.

Apart from the interpretability of the finished predictor, TPLS also offers useful information on the relative importance of different brain regions by showing the tradeoff between additional thresholding and cross validation performance (**Fig. 1-6D & E**). Because the thresholding level of TPLS model is chosen to maximize cross-validation performance, users can experiment with different thresholds to see how much predictive performance is sacrificed when fewer regions are recruited into the predictor. **Fig. 1-6F~H** shows various predictors at different thresholds where stringent thresholds (e.g., **Fig 1-6H**) highlight the more important brain regions for prediction. This analysis is possible, as discussed, thanks to TPLS's fit-once-tune-later approach which allows the user to generate and compare as many predictor maps as they want without re-fitting the model.

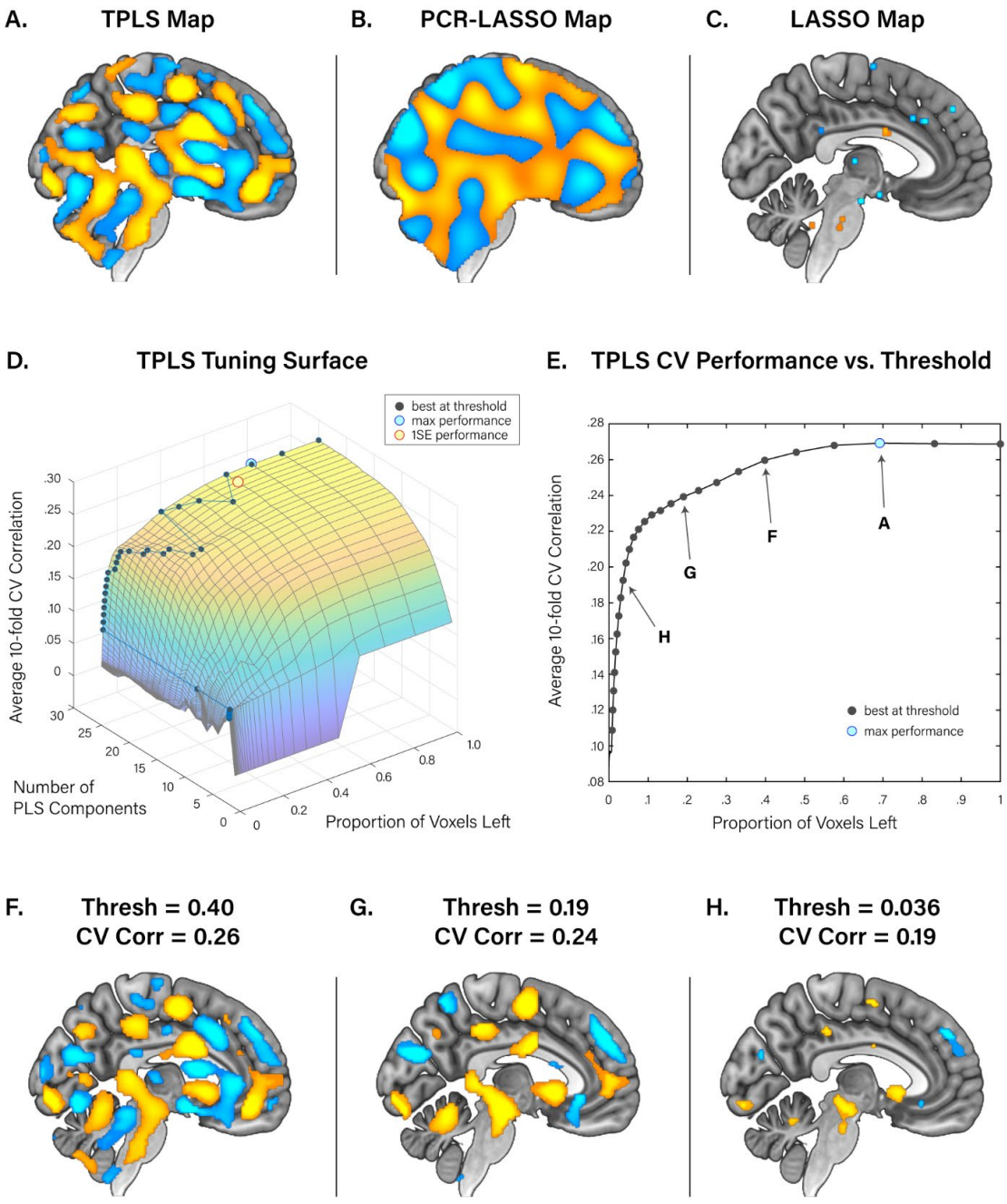


**Figure 1 - 6. Final predictors of value-based choice.**

Panels A, B, and C show the whole-brain predictor of value-based choice constructed via TPLS, PCR-LASSO, and LASSO, respectively. Panel D shows the 10-fold cross validation tuning curve for fitting the predictor. Panel E shows the 10-fold cross validation performance of TPLS model at various thresholding levels. Panel F, G, and H show the corresponding thresholded predictors from panel E.

*Simulation Results*

To compare the whole-brain prediction methods against known "ground-truths", we simulated a 2D brain signal to fit the whole-brain prediction methods (**Fig. 1-7**). We simulated a 17x17 voxel grid where only the center 5x5 grid has signal that is predictive of Y (**Fig. 1-7A**). This image was subsequently smoothed to simulate smoothness of fMRI images (**Fig. 1-7B**). When we fit a OLS predictor on this simulated data, it highlighted the canonical problem with fMRI images: multicollinearity. OLS predictors resulted in voxel coefficients that were tessellating in alternating signs, which makes it impossible to know whether the signal is positively predictive or negatively predictive (**Fig. 1-7C**). LASSO deals with this multicollinearity by selecting variables that are the most useful in prediction and removing the rest. The resulting predictor, therefore, is very sparse and is difficult to identify the region (**Fig. 1-7D**). PCR-LASSO uses PCA data-reduction to create locally smooth predictors that reflect the smoothness of fMRI images. However, because there are no voxel selection included, it can be difficult identify regions that are predictive from those that are not; this is especially detrimental in terms of not removing PCA-based artifacts which can be seen on the predictor (**Fig. 1-7E**; negative coefficients on the edges of the image). Of all these methods, TPLS provides the closest resemblance to the ground-truth signal by detecting the positively predictive 5x5 signal grid in the center, while eliminating all other voxels from the predictor. These clustered coefficients help the researcher identify the predictive regions apart from those that are not.

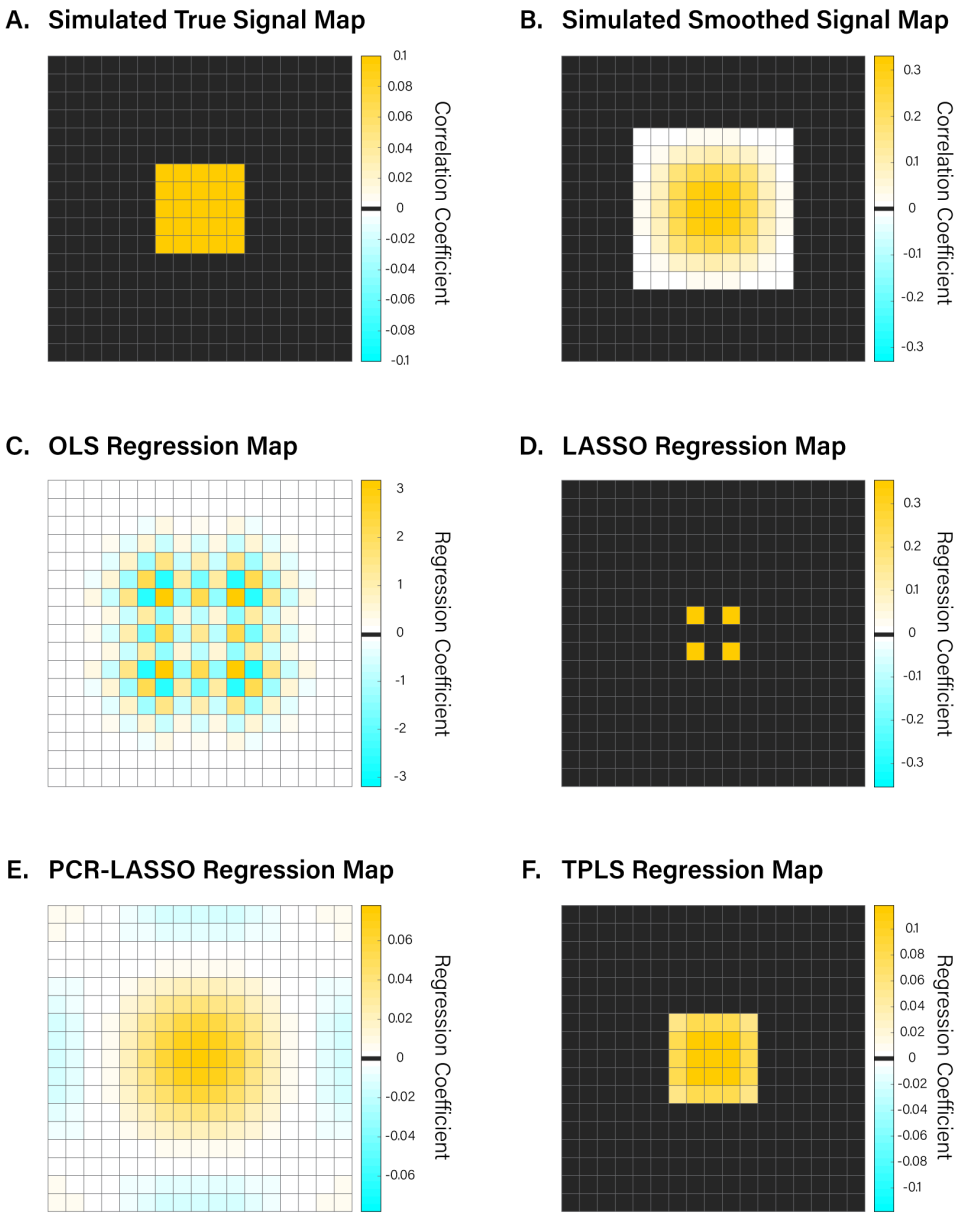


**Figure 1 - 7. Simulated neuroimaging signal and various predictor fits.**

Panels A shows a simulated 17x17 voxel grid where only the center 5x5 grid has signal that is predictive of Y at correlation $r = 0.1$. Panel B shows the result of a spatial smoothing filter to panel A, meant to simulate fMRI image smoothness. Coefficients that are exactly 0 were marked as black, while those that are close to 0 are marked as near white. Panel C~F shows regression coefficients from OLS regression, LASSO regression, PCR-LASSO regression, and TPLS, respectively. Only panels D and F have variable selection, thereby making most voxels exactly 0 (marked black).

## Discussion

With the advent of large-scale fMRI datasets, recent studies have started to develop generalized whole-brain predictors to decode mental states or to predict behaviors. Whole brain methods have great promise over traditionally used partial-brain methods as they harness more signal across the brain, identify unique neural signatures, and illuminate relationships between brain regions. However, existing methods of whole-brain predictors are limited in their usability due to computation time, memory usage, and, most importantly, interpretability. To this end, we here presented a novel method, thresholded partial least squares (TPLS), that can provide computationally feasible whole-brain predictors trained via cross validation.

TPLS exploits analytical properties of partial least squares (PLS) algorithms to dramatically reduce model fitting time, use less computational memory, and still provide high predictive performance. Compared in a real neuroimaging dataset against extant methods PCR-LASSO and LASSO, TPLS whole-brain predictors were up to 25 times faster and used as low as 20% of computation memory than existing alternative methods of whole-brain prediction. Furthermore, we found that TPLS was the only scalable method whose per-subject fitting time did not increase as the dataset size increased. TPLS method also showed high out-of-sample predictive performances that was comparable to other whole-brain predictors.

In particular, TPLS boasts a unique feature of 'fit-once-tune-later', where the model is fit only once and the two tuning parameters are chosen after the fact. This is in stark contrast to other modern prediction algorithms where models need to be re-fit for every tuning parameter combination. Not only does this allow for faster cross-validation, but it also allows researchers to explore of various tuning parameters to determine which brain regions are important for prediction and can make decisions about the best level of sparsity given the tradeoff between parsimony and performance.

We acknowledge that other existing methods such as principal component regression (PCR) and LASSO are widely used methods even outside of neuroimaging and all have a large number of statistical packages, addendums, and improvements proposed by other researchers. For example, because PCA costs huge RAM and computation time for large datasets (as shown in this paper as well), there are stochastic variants of PCA that can deal with these problems (Halko, Martinsson, & Tropp, 2011). This may reduce the computation time and RAM usage of PCR-LASSO method shown in this paper, but it still stands that the principal components may not be the best suited for prediction and that the resulting whole-brain predictor is less interpretable than TPLS methods. Also, for LASSO, there are stochastic gradient descent-based methods that can significantly reduce the model fitting time by evaluating the likelihood function on subsamples of the data. This, however, has a tradeoff with accuracy of the fit as the gradient calculations are approximations. Concordantly, we found that they were both slower and less accurate than TPLS model, and we have excluded them from this paper.

We see the relationship between whole-brain prediction methods and ROI-based prediction methods as similar to that between multivariate regression and pairwise bivariate correlation. ROI-based methods may yield insight about how one specific region is related to a mental process, but a whole-brain method can yield insight about how multiple ROIs relate to one another and contribute to the prediction; they are both important analysis tools that every researcher must use to understand the whole picture. Concordantly, we see whole-brain prediction as an important analysis tool in coming years of neuroimaging, and we hope that the method that we propose here, along with the provided packages, can make the analysis convenient and an essential part of the pipeline as multivariate regressions are.

**Supplemental Materials**

*Neuroimaging data quality control*

Some of the data from Kable et al. (2017) was removed from the analysis due to trivial reasons. To keep the number of observations per subject roughly similar, we excluded four pilot participants who had more trials than others. Counting both session 1 and session 2 data, we had 286 sessions worth of data, each with 120 binary choices. From here, we removed 4 intertemporal choice sessions and 6 risky choice sessions that had premature termination of scan due to technical issues. Additionally, 13 intertemporal choice sessions were removed for having extremely unbalanced choices (either accept or reject more than 95% of the time), and 6 sessions were removed for having too many missed responses (more than a quarter worth of session). For risky choice, 10 sessions were removed for unbalanced choices and 6 sessions were removed for too many missed responses. In total, we had 264 sessions worth of data for ITC and 267 sessions worth of data for RC. While most participants had both ITC and RC tasks, since several subjects only had one task, we decided to treat these two tasks' sessions as separate participants for our analyses.

*Image preprocessing and single trial deconvolution*

The Kable et al., (2017) dataset was acquired with a Siemens 3T Trio scanner with a 32-channel head coil. High-resolution T1-weighted anatomical images were acquired using an MPRAGE sequence (T1 = 1100ms; 160 axial slices, 0.9375 x 0.9375 x 1.000 mm; 192 x 256 matrix). T2*-weighted functional images were acquired using an EPI sequence with 3mm isotropic voxels, 64 x 64 matrix, TR = 3,000ms, TE = 25ms, 53 axial slices, 104 volumes. B0

fieldmap images were collected for distortion correction (TR = 1270ms, TE = 5 and 7.46ms). The images were preprocessed via fMRIPrep. The preprocessing pipeline, in short, performed motion-correction, slice-time correction, and b0-map unwarping on all runs and registered and resampled to a MNI 2mm template. The authors of fMRIPrep has requested the automatically generated preprocessing info to be pasted into the manuscript in its unaltered form. They are attached below in the next section (fmriprep boilerplate).

For estimating the activity of each trial, we used beta-series regression (Rissman, Gazzaley, & D'Esposito, 2004). The regressors were time-locked to the trial onset period with event duration of 0.1 seconds and convolved with a gamma HRF function. The last trial of each run was excluded from analysis because the BOLD activity of the last trial was often not observed due to the termination of the scan. This gave us 29 regressor of interest per 1 run of scan. Additionally, we included the following nuisance regressors which were generated from fmriprep: cosine components for high-pass filtering, CSF signal, white matter signal, global signal, standard 6 motion regressors, and 6 PCA components from an anatomical mask of white matter and CSF ('a_comp_cor'). After the single trial coefficients were estimated, all images were smoothed with a FWHM 5mm gaussian filter. To make analysis easy, we only used the voxels that were active for all subjects; this gave us a fairly conservative mask of the brain with 184,319 voxels.

*Fmriprep boilerplate*

Results included in this manuscript come from preprocessing performed using fMRIPrep 20.0.5 (Esteban, Markiewicz, et al. (2018); Esteban, Blair, et al. (2018); RRID:SCR_016216), which is based on Nipype 1.4.2 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

*Anatomical data preprocessing.* A total of 2 T1-weighted (T1w) images were found within the input BIDS dataset. All of them were corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al. 2010), distributed with ANTs 2.2.0 (Avants et al. 2008, RRID:SCR_004757). The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang, Brady, and Smith 2001). A T1w-reference map was computed after registration of 2 T1w images (after INU-correction) using mri_robust_template (FreeSurfer 6.0.1, Reuter, Rosas, and Fischl 2010). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym],

*Functional data preprocessing.* For each of the 18 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A B0-nonuniformity map (or fieldmap) was estimated based on a phase-difference map calculated with a dual-echo GRE (gradient-recall echo) sequence, processed with a custom workflow of SDCFlows inspired by the epidewarp.fsl script and further improvements in HCP Pipelines (Glasser et al. 2013). The

fieldmap was then co-registered to the target EPI (echo-planar imaging) reference run and converted to a displacements field map (amenable to registration tools such as ANTs) with FSL's fugue and other SDCflows tools. Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using flirt (FSL 5.0.9, Jenkinson and Smith 2001) with the boundary-based registration (Greve and Fischl 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox and Hyde 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al. 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two

CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al. 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

CHAPTER 3 – Neural Correlates of Value Are Intrinsically History Dependent

Sangil Lee, Caryn Lerman, and Joseph W. Kable

**Abstract**

A central finding in decision neuroscience is that blood oxygen level dependent (BOLD) activity in several regions, including ventral striatum and ventromedial prefrontal cortex, is correlated with the subjective value of the option being considered, and that BOLD activity in these regions can predict choices out of sample, even at the population-level. Here we show, across two different decision-making tasks in a large sample of subjects, that these BOLD value-correlates are intrinsically negatively history dependent. If the subjective value of the previous offer was high, the signal on the current trial will be lower, and vice versa. This kind of history dependency is distinct from previously described adaptation or repetition suppression effects, but instead is of the form predicted by theories of efficient coding such as time-dependent cortical normalization. In terms of practical application, since value-based choice behavior does not exhibit the same history dependence, neural prediction studies may exhibit systematic attribution errors without accounting for history effects. The data-driven, interpretable, whole-brain prediction approach we use to identify history effects also illustrates one way to adjust predictions for neural history dependency.

## Introduction

Across hundreds of studies, researchers have consistently found neural correlates of subjective value for various commodities (e.g., taste, food, money, drugs, pictures of attractive faces and places, social approval, joy out of others misfortune, viewing of a soccer goal, etc.(Breiter et al., 1997; Cloutier, Heatherton, Whalen, & Kelley, 2008; Izuma, Saito, & Sadato, 2010; Knutson, Adams, Fong, & Hommer, 2001; Kringelbach, O'Doherty, Rolls, & Andrews, 2003; McLean et al., 2009; O'Doherty  Deichmann, R, Critchley, HD, Dolan, RJ, 2002; Pegors, Kable, Chatterjee, & Epstein, 2015; H. Takahashi et al., 2009)). Meta-analysis has revealed reliable correlates of subjective value across these studies in the ventral striatum (VS) and the ventromedial prefrontal cortex (vmPFC)(Bartra et al., 2013). Such brain region overlap suggests that there may be a common neural valuation system that operates across all of the tested commodities: a domain-general signal that increases and decreases with subjective value.

Under the assumption that these regions encode subjective value, several studies have shown that activity in VS or vmPFC can predict that individual's future choices (Knutson, Scott, Wimmer, Prelec, & Loewenstein, 2007; I. Levy et al., 2011; Smith et al., 2014; Tusche, Bode, & Haynes, 2010). Subsequently, and perhaps more surprisingly, several studies have used neural signals from these regions to predict the real-world behavior of other individuals responding to the same stimuli (Berns & Moore, 2012; Falk, Berkman, & Lieberman, 2012; Genevsky & Knutson, 2015; Genevsky, Yoon, & Knutson, 2017; Karmarkar et al., 2015; Scholz et al., 2017; Venkatraman et al., 2014). These studies employed a simple approach of using average signals from VS or vmPFC, yet they predicted behavior over and above more traditional non-brain measures (Berns & Moore, 2012; Falk et al., 2012; Genevsky & Knutson, 2015; Genevsky et al., 2017; Scholz et al., 2017; Venkatraman et al., 2014). For example, Berns and Moore (2012) showed that VS activity of adolescents listening to unknown artists' songs significantly predicted

future album sales while their self-report ratings did not; Falk, Berkman & Lieberman (2012) showed that neural activity in vmPFC while watching anti-smoking ads predicted the population-level success of those ads while self-report judgments did not; Venkatraman et al. (2014) showed that activity in VS predicted market-level response to advertising beyond traditional measures; and Genevsky, Yoon, and Knutson (2017) showed that VS measurements from a small sample predicted the market-level outcome of crowdfunding projects while behavioral measures did not.

However, despite these successes in the applied domain, a basic question remains: what quantity, exactly, is being tracked by activity in regions, such as VS and vmPFC, that exhibit value correlates? Unlike theoretic utility, which is unbounded, neural signals of value are constrained by physiological limits. Hence, in order to efficiently encode values for small stakes decisions (e.g., $1 vs. $2) as well as for large stakes decisions (e.g., 1 million vs. 2 million), the brain needs to adapt its sensitivity to the currently relevant range of values. Previous studies have shown that in blocks with large value options, neural value signals, whether measured with fMRI or single neuron recordings, exhibit lower sensitivity in order to represent a wide range of values, while in blocks with only smaller value options, neural value signals exhibit higher sensitivity to optimally represent the narrow range of values (Cai & Padoa-Schioppa, 2012; Cox & Kable, 2014; Kobayashi, Pinto de Carvalho, & Schultz, 2010; Padoa-Schioppa, 2009). One way to achieve these block-wise changes in sensitivity would be for activity to adjust to recent history in a trial-by-trial manner: specifically, activity would be elevated if the previous trial's value was low and it would be depressed if the previous trial's value was high. Hence, the value signal would not only be positively correlated with the value offered on the current trial, but also negatively correlated with values offered on immediate past trials. Such history dependence is further predicted by theories such as time-dependent cortical normalization (Tymula & Glimcher, 2016). However, though there is some evidence for such trial-by-trial history dependence in

single neurons in the orbitofrontal cortex (Padoa-Schioppa, 2009), whether history dependence is a general feature of value coding, whether it is present in the BOLD signal, and how the history dependence in different brain regions compares to that in behavior is unknown.

Here, we show that neural correlates of value measured with fMRI are history dependent. Using a large imaging dataset of intertemporal choice and risky choice, we show, separately across both tasks, that neural signals that are positively correlated with the value of the offer on the current trial are also negatively correlated with the recent offer values on past trials. We also show that this history dependency cannot be explained by well-known fMRI adaptation or repetition suppression effects (Barron, Dolan, & Behrens, 2013; Grill-Spector, Henson, & Martin, 2006; Naccache & Dehaene, 2001; Segaert, Weber, de Lange, Petersson, & Hagoort, 2013). In contrast to the history dependency in neural activity, we find no evidence that choices are influenced by the recent offer values on past trials. Given this discrepancy, we then build an interpretable whole-brain predictor of choice, as a data-driven method of searching for neural signals of value without history dependency. However, in examining the structure of the whole brain predictor, we find that all positively weighted regions show history dependency and the negatively weighted regions served to subtract out these history effects. In terms of basic neuroscience, these results show that BOLD value correlates are intrinsically history dependent, in a trial-by-trial manner that can account for the range normalization observed in these signals over longer timescales. In terms of practical application, these results show that neural prediction studies that use univariate signals from regions of interest will make systematic errors without accounting for these history effects; the whole-brain multivariate methods we use here provides one way to do so.

## Methods

*Dataset*

We used a dataset from Kable et al. (2017)(Kable, Caulfield, Falcone, McConnell, Bernardo, Parthasarathi, Cooper, Ashare, Audrain-McGovern, Hornik, et al., 2017). The purpose of their study was to examine the effect of commercial cognitive training software on brain activity during decision making tasks. Participants completed two sessions of economic decision-making tasks with 10 weeks in between. In each session, they completed 4 runs each of intertemporal choice (ITC) and risky choice (RC) tasks while being scanned. Each run had 30 binary choices. In ITC, the choices were between a smaller immediate reward of $20 today and a larger later reward (e.g., $30 in 7 days); in RC, they were between a smaller certain reward of $20 and a larger probabilistic reward (e.g., $40 with 60% chance). In both tasks, the smaller amount was always $20, while the larger option varied in amount and delay/probability. The original study reported no difference in decision-making or brain activity after 10 weeks. Hence, in the current paper, we combine both session 1 and session 2 datasets. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Of the 160 participants that completed session 1, we excluded those with missing runs (n = 6), head movement (any run with >5% of mean image displacements greater than 0.5mm; n = 3), more than 3 missing trials per run for two or more runs (n = 2), or revealed study design (n = 1, and one subject expressed awareness of their experimental condition, i.e. cognitive training vs. control). Of the remaining 148, 114 also completed session 2, from which we also excluded those with missing runs (n = 3), head movement (n = 2), or too many missing trials (n = 2). This left us with 148 participants' session 1 data and 107 participants' session 2 data. To ensure the reliability of our choice models, we excluded sessions with extremely one-sided choices (less than 10 trials of either choice type). This removed 20 sessions in ITC and 14 sessions in RC. In total, our final

dataset was 235 sessions of ITC data and 241 sessions of RC data across 147 participants and 2 sessions (1 participant's data was completely removed due to one-sided choices in both tasks).

*fMRI image acquisition, preprocessing*

The brain images were collected with a Siemens 3T Trio scanner with a 32-channel head coil. T1-weighted anatomical images were acquired using an MPRAGE sequence (T1 = 1100ms, 160 axial slices, 0.9375 x 0.9375 x 1.000 mm, 192 x 256 matrix). T2*-weighted functional images were acquired with an EPI sequence with 3mm isotropic voxels (TR = 3,000ms, TE = 25ms, 53 axial slices, 64 x 64 matrix, 104 volumes). B0 images were collected for unwarping (TR = 1270ms, TE 1 = 5.0ms, TE 2 = 7.46ms). Images were preprocessed with FSL: skull stripped with BET (Brain Extraction Tool), motion corrected with MCFLIRT (Linear Image Restoration Tool with Motion Correction), spatially smoothed (FWHM 9mm Gaussian Kernel), high pass filtered (cutoff = 104s), and registered to MNI space with FNIRT (FMRIB Non-linear Image Registration Tool).

*Behavioral Analyses*

Behaviorally, we test if people's choices are affected by the subjective value of previous trials' offers using choice modeling. In each of the two tasks separately, we modeled the logit of each trial's choice probability as a weighted sum of current and previous trials' subjective values, which were simultaneously estimated with a utility function. While researchers often use parametric utility models, there is a myriad of established models for ITC and RC (e.g., exponential (Samuelson, 1937), hyperbolic (Herrnstein, 1981), generalized hyperbolic (Green,

Fry, & Myerson, 1994), quasi-hyperbolic (Laibson, 1997); expected utility theory (Morgenstern & Von Neumann, 1953), prospect theory models (Goldstein & Einhorn, 1987; Prelec, 1998; Tversky & Kahneman, 1992), mean variance models(Markowitz, 1959; Weber, Shafir, & Blais, 2004)). Furthermore, using a particular parametric formula assumes that all participants have the same utility function. Hence, instead, we fitted a data-driven utility model via cubic Bezier splines (Lee, Glaze, Bradlow, & Kable, 2019), which can approximate any of the established utility models, but can also fit unconventional forms as well. (We note, though, that using other established utility models, such as hyperbolic discounting and power utility functions, does not change our substantive conclusions.)

The model form is shown below. Let $Y_t$ denote the choice at each trial such that $Y_t = 1$ is choosing the larger reward and $Y_t = 0$ is choosing the smaller \$20 at trial $t$. We fit the following model to ITC and RC data for each session (120 choices) using MLE:

$$logit\big(p(Y_t = 1)\big) = \beta_{scale}\left(\beta_0 + \widehat{SV_t} + \sum_{i=1}^{n} \beta_i \widehat{SV_{t-i}}\right) \quad (1)$$

$$\widehat{SV_t} = Amt_t \cdot CBS(delay_t), \quad or \quad \widehat{SV_t} = Amt_t \cdot CBS(probability_t) \quad (2)$$

$CBS(delay_t)$ and $CBS(probability_t)$ shown above denote the discounting function constructed with Cubic Bezier Splines that take delay or probability as input and outputs a discounting factor between 0 and 1. The first $n$ trials of each run were removed as they have no lagged variables. Because we estimate the overall scale of the coefficients ($\beta_{scale}$), the coefficient of the current subjective value ($SV_t$) was fixed at 1. The coefficients of the lagged subjective values ($\beta_1 \sim \beta_n$) were constrained between -1 and 1. We fitted two models: lag 4 and lag 0. The lag 4 model was fitted to assess if there was any significant influence of past trials. Upon concluding that there

was not, the lag 0 model was fitted to calculate each trial's subjective value (eq. 2), which were in turn supplied to the fMRI GLM analyses (see below).

*fMRI Analyses – GLM*

Parallel to the behavioral analysis, we test if neural signals are affected by the subjective value of past trials' offers; separately for both tasks, we regressed each voxel's activity against the intercept, the subjective value (SV) of the offer on the current trial and on each of the previous four trials. To remove their effects from the regression, one nuisance regressor modeled the average activity of the first four trials (trials that have no lagged values). The lagged SV regressors were orthogonalized with respect to the intercept and previous lags (e.g., a lag2 regressor was orthogonalized with respect to the average, lag0, and lag1 regressors). All regressors were time-locked to the trial onset with event duration of 0.1 seconds and convolved with a gamma hemodynamic response function. We used FSL's FLAME (FMRIB's Local Analysis of Mixed Effects) model to obtain an average coefficient of each session for each task. Then we tested for brain regions that have significant activity at the group level by performing permutation tests across all the sessions (235 sessions for ITC, 241 sessions for RC). We used a threshold of $p = .01$ for each lag ($0 \sim 4$) so that the Bonferroni familywise error rate would be $p = .05$ for each task.

To thoroughly test whether regions where activity was correlated with the current offer's SV are also influenced by the SV of previous trials' offers, we examined the lagged regressor coefficients in three different ways. First, we created ROIs by choosing voxels from the Bartra et al. meta-analysis (Bartra et al., 2013) that have significant effect of SV in both of our tasks. Using these ROIs for vmPFC and VS, we estimated their average lagged coefficients. Second, we

defined spherical ROIs (radius = 2 voxels) around each peak for the effect of the current trial's SV and estimated the average lagged coefficients in these ROIs. Peaks were determined as the local maxima of permutation t-stats within a radius of sqrt(12) voxels (64 ROIs for ITC and 75 ROIs for RC). Last, we examined the raw distribution of lag1 and lag2 coefficients for all voxels where activity was significantly correlated with the current trial's SV at the whole-brain level from the aforementioned permutation tests.

*fMRI Analyses – Repetition Suppression*

To formally test repetition suppression, we ran 5 GLM analyses each with a different model for repetition suppression. Each GLM included an average activity regressor for all trials, a subjective value regressor, and a distance measure between current and last trial. We employed 5 different distance measures to assess all possible scenarios of repetition suppression: 1) absolute distance between current and previous trial's subjective value, 2) absolute distance between the current and previous trial's reward amount, 3) absolute distance between the current and previous trial's delay/probability, 4) "cityblock" distance between current and last trial's offer in combined attribute space (that is, amount and delay/probability, the additive combination of 2 and 3), and 5) Euclidean distance between current and last trial's offer in combined attribute space. Each trial's subjective value was calculated using the logit model specified in eq (1) and (2) with lag = 0. Each trial's attribute regressors (i.e., amount, delay/probability) were normalized to [0 1] range (amount was divided by 85, which is the maximum amount; delay was divided by 180 which was the maximum delay). For each combination of task and GLM model, we tested for significant regions of repetition suppression by permutation testing the individual coefficients at the group-level with threshold-free cluster enhancement and two-tailed tests $p < .05$.

*fMRI Analyses – MVPA*

We combined fMRI data across subjects and the two tasks to build a generalized predictor of choice. First, we estimated the fMRI activity of each single trial via the LS-S (Least Squares-Separate) method(Turner, 2010). This involved running a separate GLM for each trial with two regressors: one regressor for the trial of interest and one nuisance regressor modeling the average activity of all other trials. The last trial of each run was excluded from this analysis because the scan was often terminated before the onset of blood oxygen level dependent (BOLD) activity of this trial. To downplay the role of noisy estimates, we used the t-stats instead of the raw coefficients. The t-stats were registered to a standard MNI 3mm template before further analysis. Then, we removed voxels that had .4 or less prior probability of being grey matter (FSL's grey matter prior).

We used principle component logistic regression with bootstrap thresholding to create an interpretable whole-brain predictor (PCLR-bootstrap). PCLR-bootstrap starts with principle component decomposition(Halko et al., 2011) of the predictor matrix, which gives two matrices: the components and their loadings ($\mathbf{X} = \mathbf{TW}$, where $\mathbf{T}$ is component and $\mathbf{W}$ is loadings). The components are then used as predictors in logistic regression ($\text{logit}(\mathbf{Y}=1) = b_0 + \mathbf{T}\hat{\mathbf{b}}$). The resulting coefficients are then projected back into the brain space by multiplying them with their respective loadings ($\boldsymbol{\beta} = \mathbf{W}\hat{\mathbf{b}}$). This process was repeated 5000 times with bootstrapped trials to obtain the *p*-value of each voxel, which was used to threshold the whole-brain map. The resulting sparse whole-brain coefficient predictor can be conveniently used by simply multiplying it to any new brain image to yield a prediction score (i.e., score = $\mathbf{X}\boldsymbol{\beta}$).

In order to find the best predictive whole-brain map, we performed a grid search of two PCLR-bootstrap tuning parameters while performing cross validation: the number of principle

components (400, 600, 800, 1000, 1500, 2000, 2500), and *p*-value threshold ($10^0 \sim 10^{-2}$ in 50 log intervals). For each combination of the two parameters, we performed leave-one-subject-out cross validation across 147 participants. The predictor was trained with 146 participants' data (combined across tasks and sessions) and predicted the left-out participant's data. Out-of-sample prediction performance was measured via the area under the receiver operating characteristic curve (AUROC). The mean of these 147 AUROC scores were used to guide the selection of the best number of principle components and *p*-value. While the best predictive map was constructed with 1500 principal components with threshold at $p = 0.324$, we found that the particular choice of these two parameters were not too important since the parameter combination with the lowest LOOCV performance was only lower by 1% AUROC. For comparison, we also measured the cross-validation AUROC of Bartra et al. ROI predictors which is the mean activity within the ROI.

In order to interpret the whole-brain predictor, we lowered the *p*-value threshold to .01 so that we can focus on the most significant voxels. This led to 44 spatially distinct clusters. One way to understand the whole-brain predictor is to think of it as a sum of these 44 regions' signals, which, in turn, is the weighted sum of each region's voxels' activity according to their assigned weight maps from the whole-brain predictor. To assess the relationship between these 44 signals, we performed hierarchical clustering analysis using the correlation between the regions' signal as the distance metric. Pairwise correlation between the 44 signals were used as a distance metric (1 – Pearson *r*) and average distance between clusters were used to calculate the hierarchical clustering linkage (MATLAB *linkage* function).

**Results**

We examined three aspects of history dependencies in valuation. First, in behavioral data, we examined if choices are affected by the subjective value of the offers on previous trials, in addition to the subjective value of the current offer. Then, in the neural domain, we examined if BOLD activity in brain regions where the signal is correlated with the subjective value of the current offer is also correlated with the subjective value of previous trials' offers. As summarized in **Figure 2-1a**, we found that choices were not affected by the subjective value of the offers on previous trials, but neural activity was affected negatively by them. In light of this potential difference in history-dependence between neural signals and behavior, we built a whole-brain choice predictor to see how neural data can be used to predict behavioral choices.

Participants completed two economic decision-making tasks while being scanned (**Figure 2-1b**). In ITC, participants made binary choices between a delayed larger monetary outcome, which was shown on the screen, versus an immediate smaller monetary outcome of $20 (fixed across all trials), which was not shown on the screen. In RC, participants made binary choices between a risky larger monetary outcome, which was shown on the screen, versus a certain smaller monetary outcome of $20 (fixed across all trials), which was not shown on the screen. Participants used a button pad to indicate whether they would accept the larger (but delayed or risky) option on the screen or to reject that offer and receive the smaller (but immediate or certain) $20.

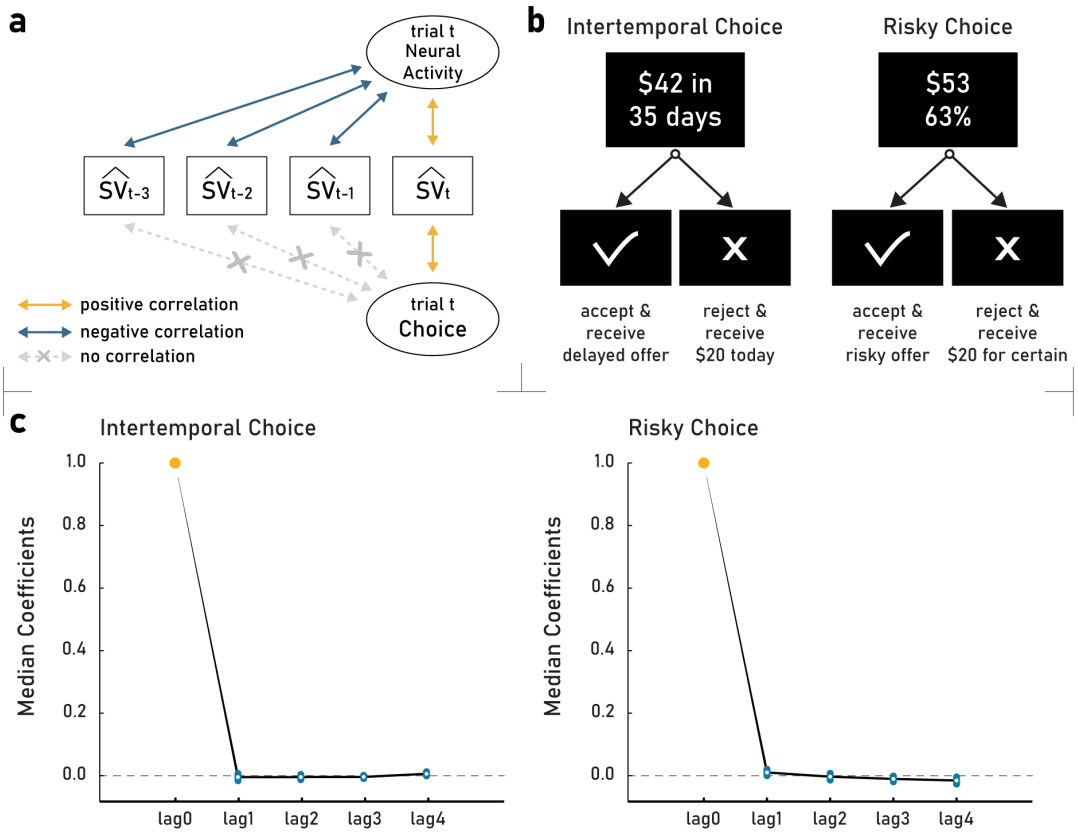


**Figure 2 - 1. Summaries of regression analyses (a), summary of task (b), and lagged subjective value regression from behavioral data (c).**

Panel (a) above shows a simplified schematic of the findings in regression analyses. We found that neural activity in value regions was not only positively correlated with subjective value estimates of the current trial's offer but also negatively correlated with subjective value estimates of the past trials' offers. On the other hand, we found no evidence that choice on the current trial was influenced by subjective value of the past trials. Panel (b) below shows a simplified schematic of the two tasks intertemporal choice (ITC) and risky choice (RC). In each task, participants saw a larger monetary option on the screen (delayed/risky) which they could either accept or reject. Panel (c) is the median coefficient estimates from a lag 4 behavioral model (eq. 1) fitted for 235 ITC sessions and 241 RC sessions. The error bars are standard errors of the median calculated via bootstrap. The coefficient of the subjective value (lag0) was fixed at 1, and was not estimated, but is plotted to show relative contribution of each variable. Two-tailed sign-rank tests were performed to test the significance at each lag. Multiple comparison was corrected at the task level using Holm-Bonferroni correction. No lags were significantly different from zero at $p < .05$ level.

*Choice is not affected by the value of previous trials' offers*

Behaviorally, we found that choices are not significantly affected by the subjective value of the previous trials' offers. We used logistic regression with lagged subjective values to see if the subjective values of the offers on the past 4 trials have any influence on the current choice. As shown in **Figure 2-1c**, there is very little influence of past trials' subjective values. To maximize power, we also fitted a simpler model with only the current and immediately previous trials' subjective values. Even then, however, the coefficient of the previous trial's subjective value was not significantly different from zero at the group level and was tightly distributed around zero (sign-rank $z$ = -0.88, $p$ = 0.38 for ITC; $z$ = 0.10, $p$ = 0.92 for RC).

*Neural correlates of value are affected by the value of previous trials' offers*

Neural analyses, on the other hand, revealed many areas where activity is significantly influenced by the subjective value of previous trials' offers (**Figure 2-2**). We found that activity in many regions is significantly positively correlated with the current trial's subjective value but also negatively correlated with the last trial's subjective value. No negative correlations were found for the current trial's value and no positive correlations were found for the past trials' value. VS and vmPFC were among the regions showing overlapping positive effects of the current trial's value and negative effects of the previous trial's value.

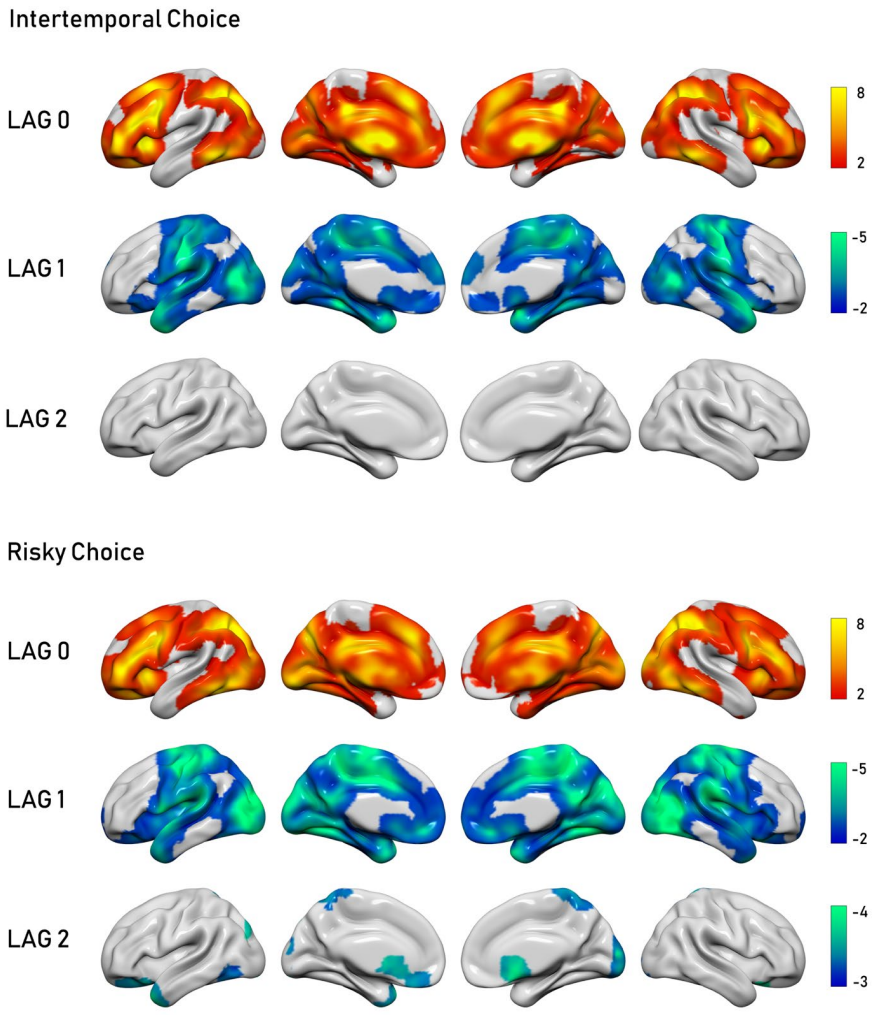


**Figure 2 - 2. t-statistics from whole-brain permutation tests for GLM lagged coefficients.** For each lag (0 ~ 4), we performed two-tailed permutation tests at p = .01 level to find regions that were significantly correlated with each lagged regressor. This corrects for multiple comparisons for each task at p = .05 (Bonferroni correction). No activity was significant beyond lag 2.

Region of interest analyses further confirm the prevalence of lagged value effects. First, we created ROIs for VS and vmPFC by selecting the voxels that showed significant subjective value effects in both of our tasks (i.e., the conjunction of lag 0 maps from **Figure 2-2**) within the regions identified by the Bartra et al. (2013) meta-analysis. We obtained the mean coefficients within each of the two ROIs and then tested if they were significantly different from zero at the

group-level (**Figure 2-3**). In both ROIs and in both tasks, we found that activity was largely negatively correlated with previous trials' values with diminishing effect as the lags increase.

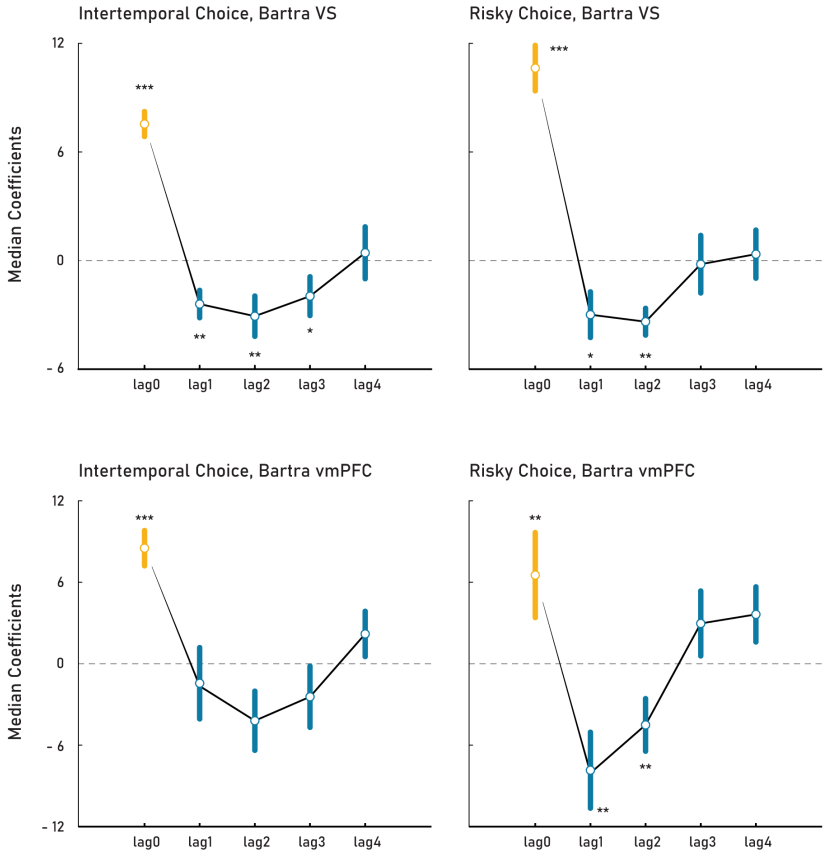


**Figure 2 - 3. Lagged subjective value coefficients from conjunction mask with Bartra et al. (2013) ROIs.**
ROIs were created by a three-way conjunction between Bartra et al. masks and permutation tested maps for subjective value in our two tasks. Average coefficient was calculated by taking the median of all coefficients within the ROI. Above plot shows the median of these coefficients across all sessions (235 ITC sessions and 241 RC sessions). The error bars are standard errors of the median calculated via bootstrap. The left panels show the lagged coefficients of ITC and the right panels show that of RC. The top panels show the results from the ventral striatum (VS) ROI and the bottom panels show that from the ventromedial prefrontal cortex (vmPFC) ROI. All significance testing was performed via two-tailed sign-rank tests. * $p < .05$, ** $p < .01$, *** $p < .001$.

We found similar results across 64 ROIs in ITC and 75 ROIs in RC defined based on the peak positive effects of the current trial's subjective value (**Figure 2-4**). Across both tasks, an overwhelming majority of the ROIs had negative lag1 and lag2 coefficients, suggesting that

regions with a positive effect of subjective value on the current trial tend to have a negative history effect. We also corroborated these results at the voxel level by finding that most of the voxels showing a significant correlation with subjective value on the current trial have negative lag 1 and lag 2 coefficients (**Figure 2-4**).

*Neural history effects are distinct from repetition suppression*

Previous fMRI studies have identified repetition suppression effects – reduced activation for repeated stimuli – across a wide variety of different stimuli and brain regions, including in vmPFC. The history dependency we observe in neural value correlates is similar to repetition suppression, in that both involve an influence of the previous trial's stimulus on the current trial's neural response. However, the history dependency described above differs from repetition suppression in that it is independent of the current trial's stimulus, while repetition suppression depends critically on the similarity between the current and previous trial's stimulus. For example, the history dependency described above would predict the activity on the current trial is reduced when the previous trial's offer value is high, regardless of the current trial's offer, while repetition suppression would predict that activity on the current trial is reduced when the previous trial's offer value is high, if the current trial's offer value is also high. Given that our trials are randomly ordered, the similarity or distance between the current and last trial's offer should be uncorrelated with the past trial's offer, and thus the two effects should be distinguishable in our data.

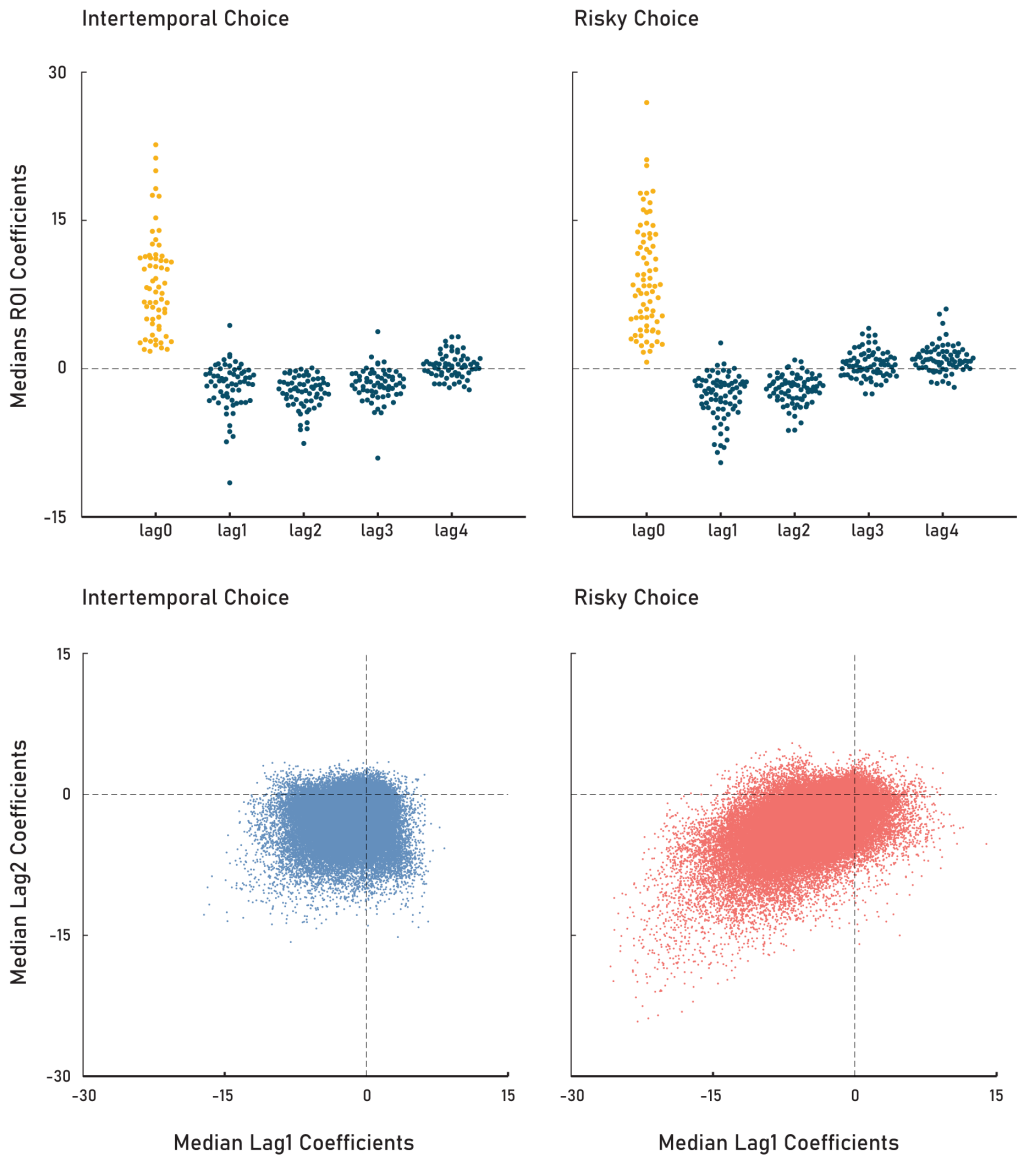


**Figure 2 - 4. Distribution of lagged coefficients for SV peak ROIs and SV voxels.**

There were 64 ROIs for ITC and 75 ROIs for RC (each represented as a dot on the upper panel) that were determined by constructing a spherical ROI around the peak voxels of lag 0 t-stat map from figure 3 (i.e., voxels that were shown to be significantly positively correlated with current trial's value at the whole-brain level). The mean coefficient within each ROI was taken as an average measure for each session, and the median of these measures across all sessions are plotted above. Bottom panel is Scatterplot of median lag 1 and lag 2 coefficients for voxels that show significant correlation with subjective value of current (lag0) trial. Each dot is a single voxel whose coordinates are medians of the lag 1 and lag 2 coefficients across all participants.

To show that the history dependency we describe above cannot be accounted for by repetition suppression, we systematically examined five different forms of similarity or distance between the current and previous trials' offers, involving the subjective value of the offers, the attribute values (amount or delay/probability) of the offers, or combinations of the attribute values (assuming additive/"cityblock" or Euclidean distance in two-dimensional attribute space). We found no significant effects of value distance in the brain, and the only distance effects that were consistent across the two tasks were for combined attribute distance in the intraparietal sulcus (**Figure 2-5**). Critically, none of the distance effects were significant in VS or vmPFC. Even when we used the Bartra ROIs for VS and vmPFC, none of the regions showed significant positive activities for any of the 5 different regressors across both ITC and RC.

*Whole-brain neural predictor cancels out the history signal in neural data*

Given the discrepancy between strong history effects in neural data but no history effect in behavioral data, we used a data-driven approach to search for any history-independent value signals in the brain. Since choice has no history effects, building a whole-brain predictor of choice should capitalize on any neural signals of value without history effects, if such signals exist. Through a PCLR-bootstrap procedure (**Figure 2-6**), we created a whole-brain predictor of choice with LOOCV performance (as measured by area under the ROC curve: AUROC) that exceeded that of simple mask-based approaches to prediction.

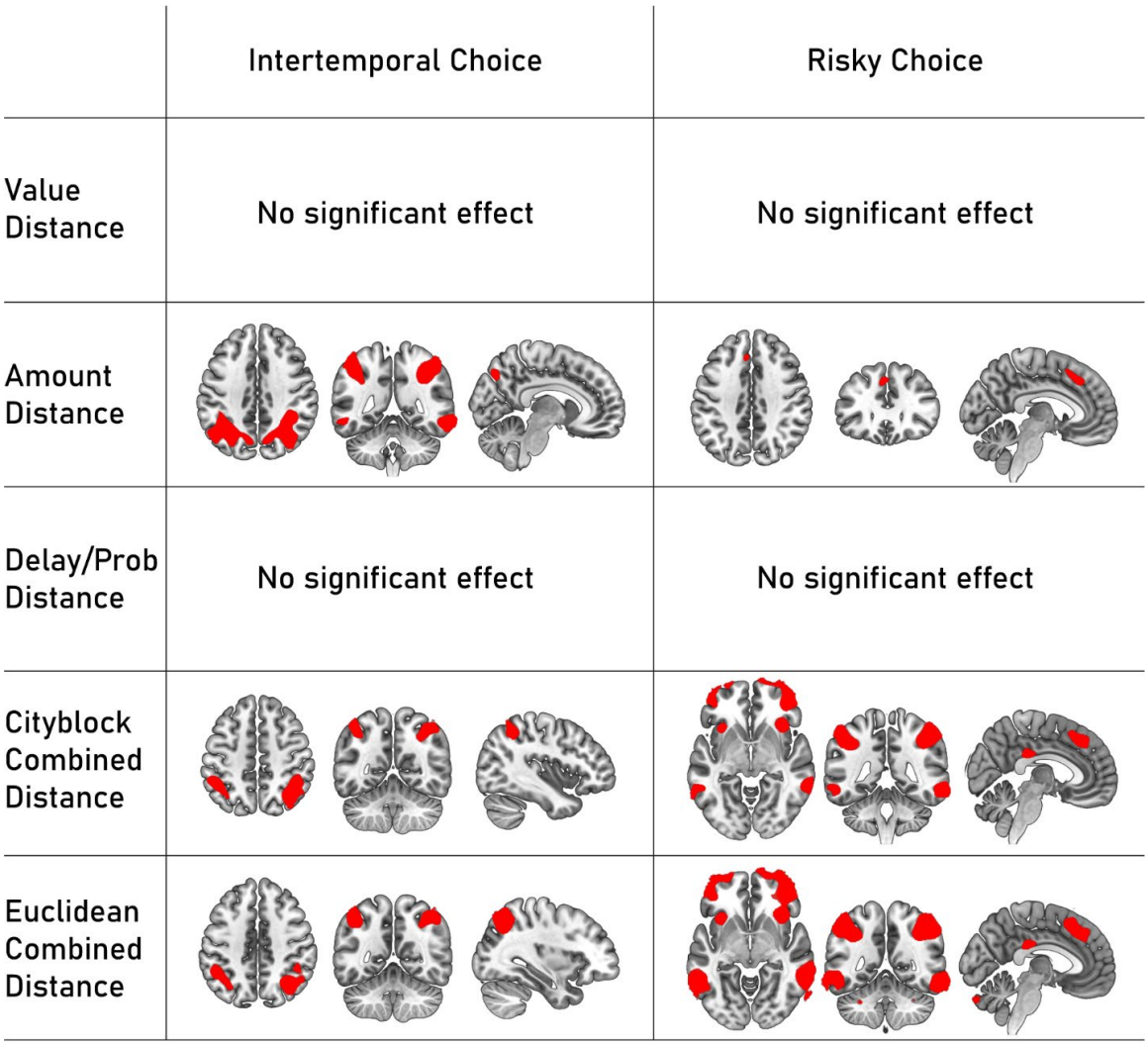


**Figure 2 - 5. Significant regions for repetition suppression regressors in the brain.**

Each row corresponds to a different model of repetition suppression. The left column shows significant regions for intertemporal choice dataset and the right column shows significant regions for risky choice dataset. Result for each cell is tested with permutation testing with threshold-free cluster enhancement two-tailed t-tests at $p < .05$. All found results were positive effects. No negative effects were found.

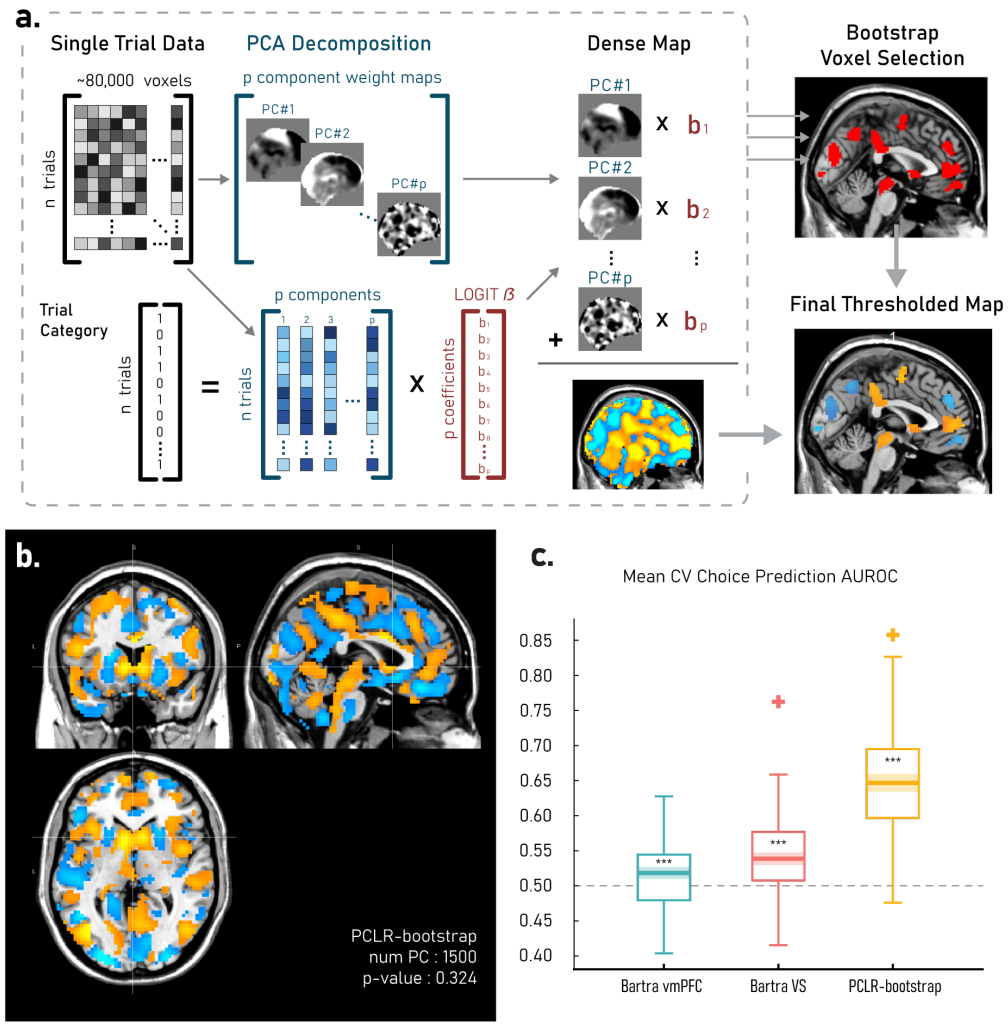


**Figure 2 - 6. Flowchart of PCLR-bootstrap (a), resulting whole-brain predictor (b), and cross-validation performance (c).**

Top panel (a) shows the flowchart of PCLR-bootstrap. The original data are composed of single-trial images of the brain and their corresponding behaviors (choices). The single-trial images are then decomposed into principle components, which in turn are used as predictors in a logistic regression. The coefficients of these components, when multiplied with their respective loadings, yield a whole-brain map. The logistic regression is repeated with bootstrapped samples in order to calculate a bootstrap p-value for each voxel, which are then used to threshold the whole-brain prediction map. Panel (b) shows the whole-brain predictor created with PCLR-bootstrap, and panel (c) shows its prediction performance compared with mask-based approaches. All three prediction methods yielded performance significantly higher than 50% (two-tailed sign-rank $p < .001$). The solid line in the middle of the box shows the median, with the shaded thicker line showing 95% confidence interval of the median. The boxes show the interquartile range and the whiskers show the range of data (or 1.5x the interquartile range in case of outliers). The cross mark shows the outliers that are more than 1.5 times the interquartile range away from the nearest interquartile.

A hierarchical clustering analysis of the whole-brain predictor (with more stringent thresholding at $p < .01$) revealed two distinct groups of regions that contribute unique signals to the whole-brain predictor (**Figure 2-7**). We obtained the contribution of each of the 44 contiguous regions to the prediction by calculating the weighted sum of the voxels according to the whole-brain predictor coefficients on every trial. Then the pairwise correlation between these 44 signals across all trials was used to calculate a hierarchical clustering solution. The cophenetic correlation coefficient (ranging from 0 to 1) that measures the quality of the clustering result was very high at 0.98. Furthermore, the first division into two clusters coincides perfectly with whether a region was assigned a negative or a positive weight in the whole-brain predictor. This result could not be achieved if the positive and negative regions were both encoding subjective value but with different signs, as multiplying the negative coefficients to negative subjective value would have made it positively correlated with subjective value. This result suggests that the whole-brain predictor can be thought of as the difference between two signals: the signal carried by the positively weighted regions minus the signal carried by the negatively-weighted regions. VS and vmPFC were in the positively weighted group.

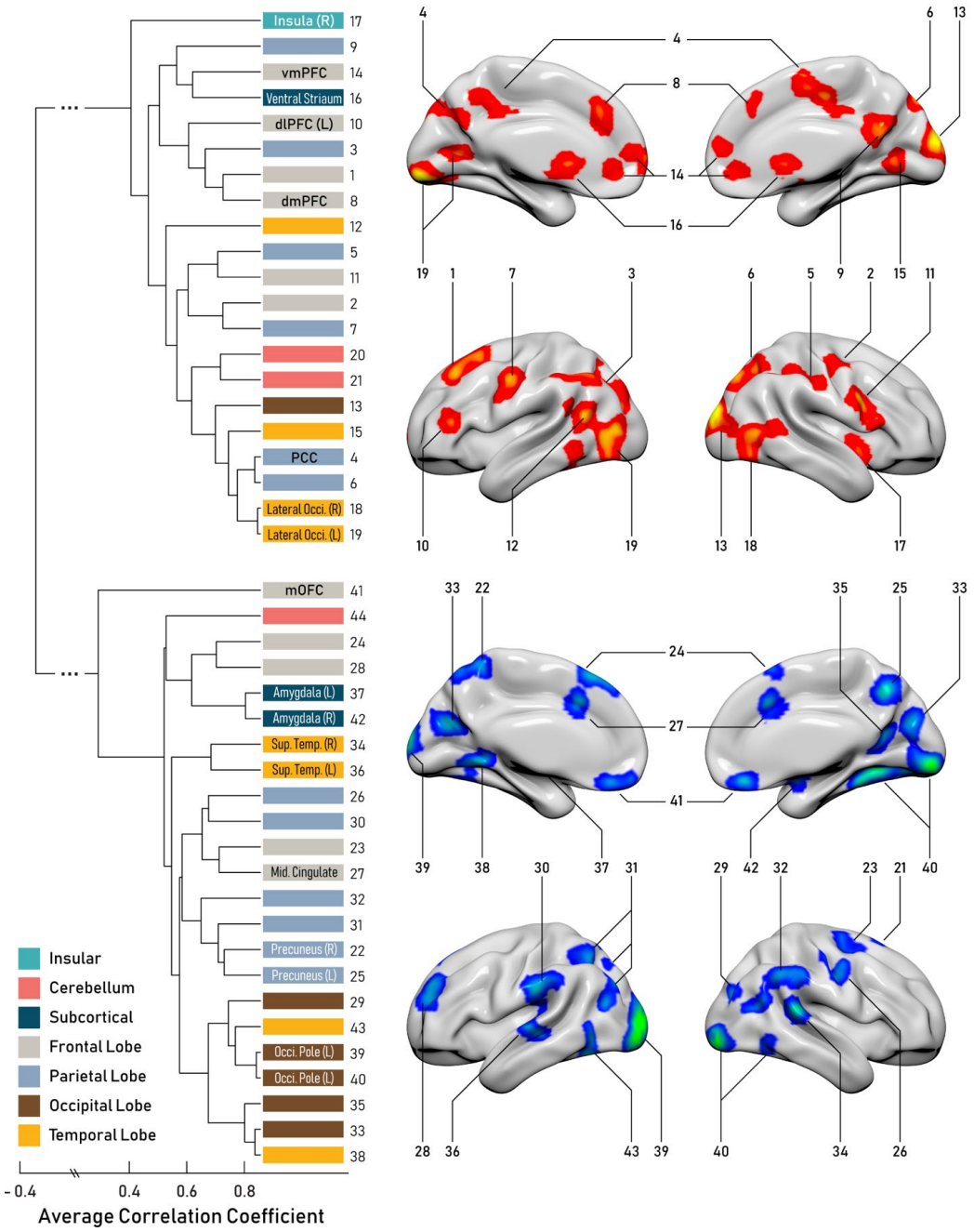


**Figure 2 - 7. Hierarchical clustering analysis of whole-brain predictor thresholded at p < .01.**

Shown on the left is a dendrogram that shows the clustering solution and marks the average correlation between groups by the height at which the branch bifurcates. The right shows the corresponding regions in each of the two groups. All positive regions ended up in one group (above) while all negative regions ended up in the other (below).

Further examination revealed that the whole-brain predictor works by using the signal in the negatively-weighted group to subtract out the history effects in the signal of the positively-weighted group. To see how the signals are being combined, we examined the average lagged coefficients of the positively-weighted group, the negatively-weighted group, and the two groups combined (**Figure 2-8**). We found that the signals from the positively-weighted group were quite similar to the signals from the negatively-weighted group; the only difference was relative strength of the subjective value signal compared to the history signal. In the positively-weighted group, the subjective value of the current offer was the strongest signal, whereas in the negatively-weighted group, the subjective value signal from the current trial was about the same size as the lagged signal from one trial back or even smaller. Due to these differences, when these signals were combined, the whole-brain predictor effectively cancels out the history effects while the subjective value signal from the current trial remains. This allows the whole-brain predictor to match the history dependency of behavior, in which we found no significant effect of the subjective value of past trial's offers on choice. The fact that the whole brain predictor is structured to subtract out history effects is striking; it suggests that such history effects are intrinsic to all brain areas that respond to value, as otherwise the whole brain predictor could have used just those regions without any effects of past history.

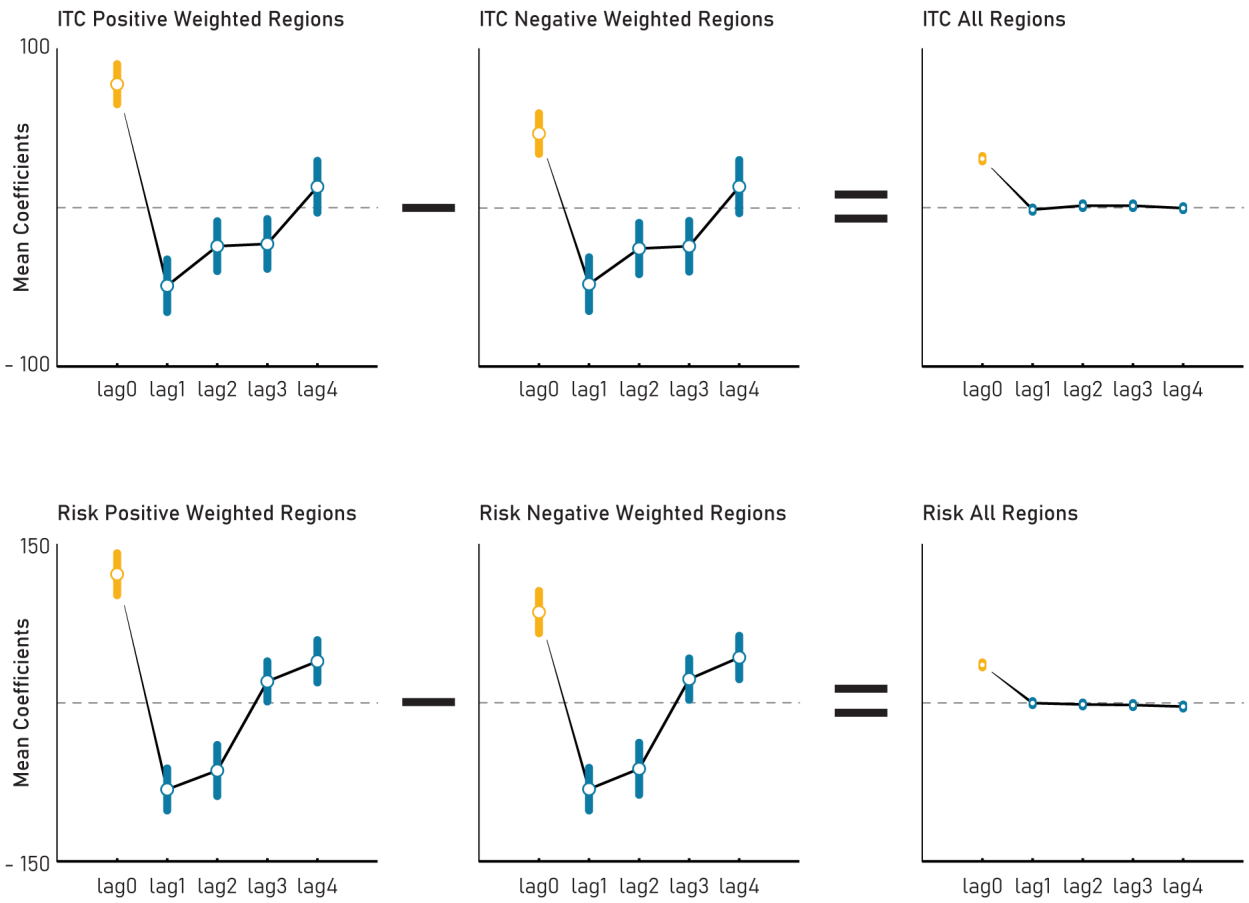


**Figure 2 - 8. Average history effects of positive regions, negative regions, and all regions of whole-brain predictor.**

For ITC (above) and RC (below) the whole-brain predictor creates a history-free signal by canceling out the history effect between the regions.

## Discussion

In the current paper, we present evidence that neural correlates of value are intrinsically history dependent. Separately across two economic decision-making tasks, we find neural signals that are not only positively correlated with the subjective value of the current trial's offer, but also negatively correlated with the subjective values of offers on immediate past trials. On the other hand, choice behavior in both tasks did not show any influence of past trial's offers. We construct

a whole-brain predictor of choice, as a data-driven way to search for neural value signals without history effects. However, examination of our whole-brain predictor shows that it does not operate by finding history-free value signals, but rather uses signals from other brain regions to subtract out the history effects intrinsic to value correlates.

Our results have both theoretical and practical implications. On the theoretical side, this is the first evidence for trial-by-trial history dependence in the correlates of value in BOLD fMRI signals. While similar trial-by-trial dynamics have previously been demonstrated in the firing rate of single orbitofrontal neurons (Cox & Kable, 2014), our results strongly suggest history dependence is a general feature of value responses across the brain, at least those observed in the BOLD signal. These trial-by-trial history dependencies could account for the range adaptation observed at longer timescales in the neural value signals measured with either fMRI or single neuron recording (Cai & Padoa-Schioppa, 2012; Cox & Kable, 2014; Kobayashi et al., 2010; Padoa-Schioppa, 2009). Both range adaptation and the trial-by-trial history dependencies observed here may be a result of efficient coding principles, which can favor encoding prediction errors to reduce redundancies in the signal(Rao & Ballard, 1999), and has recently been proposed to lead to a time-dependent cortical normalization mechanism that impacts value coding in the brain (Khaw, Glimcher, & Louie, 2017; Tymula & Glimcher, 2016; Yamada, Louie, Tymula, & Glimcher, 2018).

Critically, the kind of history dependency we observe for value correlates is distinct from the kind of repetition suppression or similarity/distance effects that have been widely studied in fMRI across different kinds of stimuli and brain regions (Barron et al., 2013; Grill-Spector et al., 2006; Naccache & Dehaene, 2001; Segaert et al., 2013). In fact, we found no significant carryover effects of value distance in the current dataset for either decision task. The distance effects that we did observe in both tasks were for the numerically presented attribute values in the

intraparietal sulcus. The anatomical location of these effects is consistent with previous studies reporting repetition suppression in parietal cortex for numerical representations (Naccache & Dehaene, 2001).

We did not observe any effects of the subjective value of previous trials' offers on the current trial's choice. This is consistent with a recent proposal that value (or other variables) is encoded according to efficient principles, and that downstream decoding and choice processes assume (and therefore adjust for) such efficient coding (Polanía, Woodford, & Ruff, 2019; Wei & Stocker, 2015). However, at least one recent study has shown that people's evaluations adapt to the range of offers across longer timescales (blocks of 90 trials) (Khaw et al., 2017). Future work should examine whether these differences across studies arise from differences in timescale, decision-making task (choices versus bids), or other factors.

Our whole-brain prediction results also illustrate how building neural classifiers or predictors with an eye towards interpretability can lead to insights into neural mechanisms. Our PCLR-bootstrap method joins other methods that emphasize interpretability (Chang, Gianaros, Manuck, Krishnan, & Wager, 2015; Grosenick et al., 2008). The clustering analysis we performed on the classifier components revealed a central feature of choice-predictive activity – intrinsic history dependence – and how the classifier compensated for this source of "noise".

On the practical side, our results suggest that neural prediction studies that use average activity from ROIs are likely to experience systematic errors unless the history effects are accounted for. In the current study, the whole brain predictor that automatically adjusted for neural history effects predicted choices with significantly higher predictive accuracy out of sample than ROI-based prediction approaches. Using ROI-based approaches, several previous studies have shown that neural measurements from a relatively small number of participants can

predict population-level behavior above and beyond traditional behavioral measures such as self-reports (Berns & Moore, 2012; Genevsky et al., 2017; Karmarkar et al., 2015; Knutson et al., 2007; Scholz et al., 2017; Venkatraman et al., 2014). We believe that using whole-brain predictors, like the one used in the current paper, will increase the predictive power of neural measurements in such applications, in part by canceling out the intrinsic history dependence of neural value signals.

# CHAPTER 4 – The Future is Less Concrete than Now: A Neural Signature of the Concreteness of Prospective Thought Is Modulated by Temporal Proximity during Intertemporal Decision-Making

Sangil Lee, Trishala Parthasarathi, Nicole Cooper, Gal Zauberman, Caryn Lerman, and Joseph W. Kable

**Abstract**

Why do people discount future rewards? Multiple theories in psychology argue that future events are imagined less concretely than immediate events, thereby diminishing their perceived value. Here we provide neuroscientific evidence for this idea. First, we construct a neural signature of the concreteness of prospective thought, using an fMRI dataset where the concreteness of imagined future events is orthogonal to their valence by design. Second, we apply this neural signature in two additional fMRI datasets, each using a different delay discounting task (bidding versus choice), to show that neural measures of concreteness decline as rewards are delayed farther into the future.

## Introduction

Many of the most important choices we make in our daily lives involve tradeoffs between the present and future. Should you spend money now or to save it for retirement? Can I endure the immediate pain of nicotine withdrawal in order to enjoy better future health of being a non-smoker? In such intertemporal decisions, humans tend to devalue, or discount, outcomes in the future; a phenomenon known as delay discounting. In the laboratory, this tendency can be measured by presenting participants with choices between a smaller monetary amount available immediately or a larger monetary amount available after a delay. Patience as measured by laboratory intertemporal choice tasks predicts other important aspects of life such as drug and alcohol abuse, educational attainment, and personal finances (Alessi & Petry, 2003; Anderson & Mellor, 2008; Brañas-Garza, Georgantzís, & Guillén, 2007; Kirby, Petry, Nancy, & Bickel, Warren, 1999; Krain et al., 2008; Lejuez, Aklin, Bornovalova, & Moolchan, 2005; Lejuez et al., 2003; Schepis, McFetridge, Chaplin, Sinha, & Krishnan-Sarin, 2011; Shamosh & Gray, 2008).

Why, however, are delayed outcomes fundamentally less desirable? Psychologists have long pondered this important question. Several theories suggest that one potential explanation is that future outcomes are less concrete. Rick and Lowenstein (2008) have pointed out that in many intertemporal decisions, delayed outcomes are intrinsically less tangible than sooner ones. For example, while smoking now has immediately perceivable pleasure for the smoker, the promise of better future health is less appreciable. Similarly, construal level theory proposes that even when future outcomes are not intrinsically less tangible, people tend to use a process of high-level construal when thinking about future events that leads to them being represented in a more abstract way (Liberman & Trope, 2014; Trope & Liberman, 2010). In contrast, when people consider sooner events, they use low-level construal and represent them in a more concrete manner. Many behavioral studies have provided some support for the central claim that the same

outcome is perceived less concretely when it occurs farther in the future rather than more immediately (for review, see Liberman & Trope, 2014). However, an ideal test of the idea that these differences contribute to discounting would measure concreteness in real time (and more directly), while people are making intertemporal decisions, and non-obtrusively, without asking the participants about concreteness directly.

Functional brain imaging has the potential to provide just such a non-obtrusive real time test of whether future outcomes are perceived as less concrete during intertemporal decision-making. Yet, while many fMRI studies have compared brain activity for sooner versus later outcomes (for review, see Carter, Meyer, & Huettel, 2010), attributing any neural differences specifically to concreteness requires ruling out other potential sources for these differences. Perhaps the most important and obvious difference between sooner and later outcomes that could drive neural activity is that sooner outcomes are valued more highly than delayed ones; that is, brain activity selectively responding for sooner versus later outcomes may reflect valuation, not necessarily concreteness. Indeed several previous studies that have compared sooner and later outcomes have found increased activity in the medial prefrontal cortex (mPFC) and posterior cingulate cortex (PCC), two regions with well-established roles in valuation (Bartra et al., 2013; Cooper, Kable, Kim, & Zauberman, 2013; Lee, Parthasarathi, & Kable, 2020; Mitchell, Schirmer, Ames, & Gilbert, 2011; Tamir & Mitchell, 2011). Instead, what is required is a neural signature that is specifically predictive of concrete versus abstract prospective thought, and independent of positive versus negative evaluation.

In the current study, we use a recently developed novel method to construct a whole-brain multivariate neural predictor of the concreteness of imagined future events (study 1). To train this predictor, we use a prospective imagination dataset (Lee et al., 2020), in which the concreteness (high versus low) and valence (positive versus negative) of imagined future events

were orthogonal, such that the subsequent neural predictor was specific to concreteness and not valence. We then applied the whole-brain neural predictor of concreteness in two separate delay discounting task datasets with different evaluation schemes (bidding vs. choice) to test whether the temporal distance of monetary options during intertemporal decision-making modulates the neural signature of concrete versus abstract imagination.

## Methods

### *Prospection Dataset*

We used a dataset from (Lee et al., 2020) to develop a whole brain predictor of the concreteness of imagined future events. This study examined neural activity associated with the valence (positive versus negative) and concreteness (high versus low) of imagined future events. Twenty-four participants underwent fMRI scanning while imagining thirty-two different future scenarios. In a 2x2 design (positive versus negative valence crossed with high versus low concreteness), eight different unique scenarios were selected for each condition based on pilot testing. Each scenario was repeated twice during the experiment. Participants completed four runs and imagined sixteen scenarios per run. Each trial involved up to 5 seconds of participants reading the scenario cue, 12 seconds of imagination, and up to 14 seconds in which participants rated the concreteness and valence of the imagined event on a 7-point Likert Scale (7 seconds each). The trial duration was buffered such that the time the participants did not use in the cue phase and the rating phase was appended to the ITI at the end of the trial to make the total duration of a single trial 34 seconds.

*Delay Discounting Datasets*

We applied the neural predictor of concreteness developed in the prospection dataset in two different delay discounting datasets to test whether the neural signature of concreteness is modulated by delay during intertemporal decisions. We use one bidding dataset and one choice dataset to evaluate the robustness of the results to different task structures. The first dataset we used was from Cooper et al., 2013, which involved bidding on delayed rewards. A total of forty participants were asked on each trial to indicate an immediate monetary amount that they would feel was equivalent to receiving $75 after a given delay, varying from 14 to 364 days. Each trial began with a screen of the form "I feel indifferent between receiving $75 in 28 days and receiving _____ now". After the prompt was shown for 3 to 5 seconds, participants were then allowed a maximum of 10 seconds to use a button pad to indicate their immediate equivalent amount within a range of $0 to $75. Each participant went through four scan runs, each of which involved twenty-six questions at different delays, ranging from 14 to 364 days. We removed one participant who bid $75 for all trials regardless of delay, as we were not sure whether the participant understood the task. An advantage of this dataset is that it presents participants with the exact same reward amount at varying delays, thereby allowing us to test whether the neural signature of imagination concreteness is modulated by the delay. The flipside of this advantage is that only the delay, and not the amount, of the delayed reward is varied across trials. This limitation is addressed in the second dataset below.

The second dataset we used was from Kable et al. 2017, which investigated the effects of cognitive training on neural activity during economic decision-making. Here we use the data from the intertemporal choice task in the first, baseline, scanning session. One hundred sixty-six participants completed four runs of the intertemporal choice task while being scanned. Each run consisted of thirty binary choices between a smaller immediate reward of $20 today that was held

constant throughout the entire session and a larger delayed reward (e.g., $30 in 7 days) that varied in amount and delay from trial to trial. On each trial, the delayed option was shown on the screen; the immediate option was not displayed. Participants pressed the left/right buttons on a button pad to indicate whether they would like to accept the delayed option shown on the screen and forego the immediate reward of $20, or to reject the delayed option and take the immediate reward of $20. Participants had up to 4 seconds to respond, and after their response, a checkmark was shown on the screen if they accepted the delayed reward and an X was shown on the screen if they rejected it.

*Image acquisition*

For all datasets, the images were collected with a Siemens 3T Trio scanner with a 32-channel head coil. High-resolution T1-weighted anatomical images were acquired using an MPRAGE sequence (T1 = 1100ms; 160 axial slices, 0.9375 x 0.9375 x 1.000 mm; 192 x 256 matrix). T2*-weighted functional images were acquired using an EPI sequence with 3mm isotropic voxels, 64 x 64 matrix, TR = 3,000ms, TE = 25ms (TE = 30ms for Cooper et al., 2013). The prospection dataset's EPI sequence involved 44 axial slices with 181 volumes (Lee et al., 2020), the intertemporal bidding dataset's EPI sequence involved 45 axial slices with 150-152 volumes (Cooper et al. 2013), and the intertemporal choice dataset's EPI sequence involved 53 axial slices with 104 volumes (Kable et al. 2017). Lee et al., (2020) and Kable et al. (2017) collected B0 fieldmap images for distortion correction (TR = 1000ms, TE = 2.69 and 5.27ms for prospection dataset and TR = 1270ms, TE = 5 and 7.46ms for the intertemporal choice dataset).

*Image preprocessing*

All datasets were preprocessed via fMRIPrep (Esteban et al., 2019). The details on the preprocessing pipeline, as generated by fMRIPrep and unaltered, are available in the supplemental materials. In short, all BOLD runs were motion-corrected, slice-time corrected, b0-map unwarped, registered and resampled to a MNI 2mm template. fMRIPrep does not perform smoothing, so it was manually performed after estimating single trial activities (see below).

*BOLD deconvolution*

We used beta-series regression (Rissman et al., 2004) to estimate the BOLD activity associated with each trial in each of the three datasets. In the prospection dataset, we estimated the BOLD activity during the imagination period of 12 seconds. The regressors were time-locked to imagination time onset with an event duration of 12 seconds and convolved with a double gamma HRF function. In the intertemporal bidding dataset from Cooper et al. (2013), the regressors were time-locked to the question period (when participants can see the prompt but cannot respond yet) with event duration of 0.1 seconds and convolved with a double gamma HRF function. Finally, in the intertemporal choice dataset from Kable et al. (2017), the regressors were time-locked to the trial onset period with event duration of 0.1 seconds and convolved with a gamma HRF function. In this dataset only, the last trial of each run was excluded from analysis because the BOLD activity of the last trial was often not observed due to the termination of the scan. After the single trial coefficients were estimated, all images were smoothed with a FWHM 5mm gaussian filter.

*Concreteness prediction map*

To create a whole brain predictor of the concreteness of imagined future events, we used thresholded partial least squares (TPLS; study 1). TPLS is similar in approach to other methods for constructing whole predictors that use principal components analysis (PCA) to reduce the dimensionality of the data followed by regression (Chang et al., 2015; Wager, Atlas, Leotti, & Rilling, 2011; Wager et al., 2013). The key advantage of TPLS over PCA-based methods is that partial least squares (PLS) is used for data-reduction. PLS components maximally explain the covariance between the predictors and the outcome, whereas PCA components only explain the variance of the predictors. Thus, PLS yields data-reduction that is more pertinent to prediction.

We built the whole-brain predictor of concreteness in three steps (**Fig. 3-1**). First, we performed PLS to extract components that maximally explain the covariance between the single trial images and the binary concreteness trial categories (high versus low). These components consist of a map of weights for each voxel in the brain. PLS also automatically yields coefficients for each component that are equivalent to the regression coefficients one would obtain from regressing the dependent variables on the components. We also calculated the t-statistics of each component as one would get from a regression model (here, given the large number of observations, we assume that the t-statistics are approximations of z-statistics). In the second step, we back-project the PLS coefficients and z-statistics into the original voxel space by multiplying them with the PLS weight maps. This yields coefficients for each of the brain voxels for easier interpretation. In the final step, we used the back-projected z-statistics of each voxel to rank their variable importance and threshold the voxel coefficient map so that less important voxels are removed from the final predictor. This final predictor can be used to obtain a 'concreteness score' for each brain image by calculating the dot product between the predictor and the image. We chose the number of PLS components to use and the level of thresholding based on the

combination that gave the highest 24-fold leave-one-person-out cross validation prediction performance as measured by the area under the receiver operating characteristic curve (AUC).

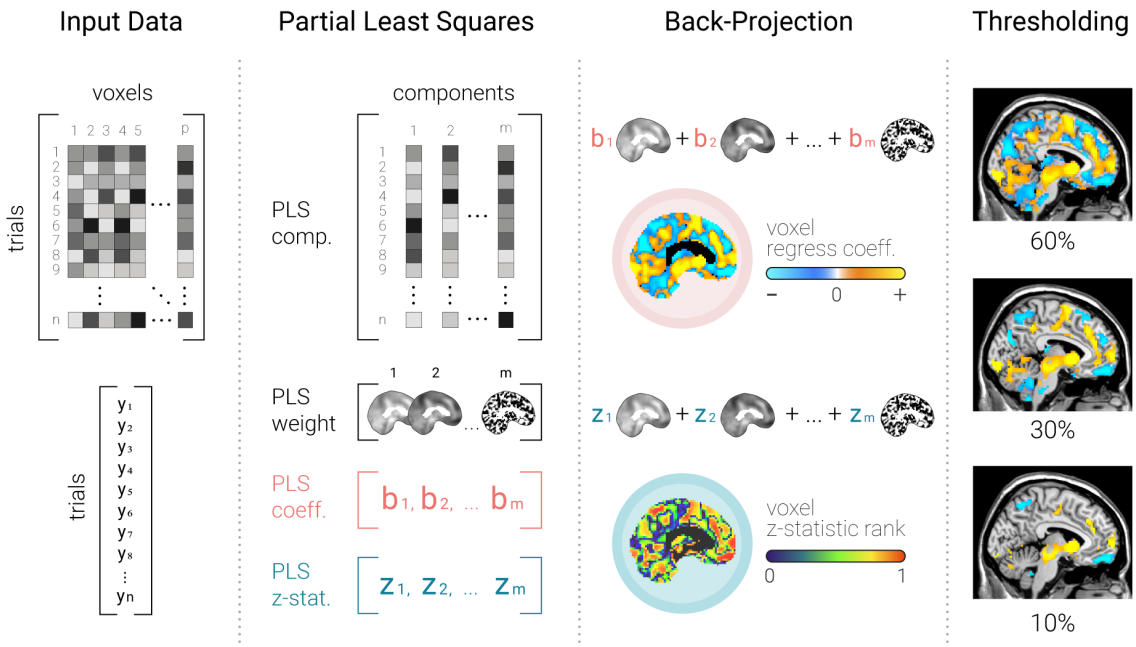


**Figure 3 - 1. Thresholded Partial Least Square (TPLS) approach to building a whole-brain predictor.**

From left to right, the TPLS method is outlined. The first step performs partial least squares on the brain image data (X) and the dependent variable (Y) in order to extract components that maximally explain the variance between X and Y. Each of these components are paired with weight maps that describe how each component is a weighted sum of the original voxels. They are also associated with regression coefficients and t-statistics (approx. z-stat) from regressing the dependent variable onto the components. These regression coefficients and z-stats are multiplied with their respective weight maps to yield regression coefficients and z-stats in the original voxel space. Using the voxel-level z-stats, the whole-brain predictor is thresholded by removing less important voxels (i.e., voxels with smaller absolute z-stats).

*Sensitivity and specificity analysis*

To assess the accuracy of our whole-brain predictor of concreteness, we performed a nested 24-fold leave-one-person-out cross validation within the prospection dataset. We trained the predictor on data from 23 participants and tested on the one left-out person. Within the 23

training participants' data, we employed an additional 23-fold leave-one-person-out cross validation to find the optimal number of components and thresholding level. After the best parameters were found, the TPLS model was fitted using all 23 participants and used to predict the left out person's data. Specifically, we tested whether the TPLS model can accurately classify the high vs. low concreteness trial categories. Furthermore, we also tested if the TPLS model predictions are correlated with the participants' self-reported ratings of concreteness.

As we trained the TPLS model on the condition labels for concreteness, and these were orthogonal by design to the condition labels for valence, we expected our whole brain predictor to be specific to concreteness and not valence. To assess the specificity or our whole brain predictor of concreteness, we also tested whether the TPLS model could not accurately classify the positive vs. negative valence trial categories, and whether its predictions are not correlated with the participants' ratings of valence.

*Concreteness and Delay Discounting*

To calculate an expression score for the neural signature of concreteness during delay discounting, we calculated the dot product between the neural predictor of concreteness and the brain image of estimated activity for each trial. These scores were then correlated with the delay until the receipt of the delayed reward (in days), and the delayed amount (for Kable et al. 2017 only, since the delayed amount is constant in Cooper et al. 2013). The correlations were performed at the individual level, and each individual's correlation coefficient was used as a summary statistic to test if there was a significant correlation at the group level.

**Results**

We first developed a whole-brain neural predictor of the concreteness of prospective thought. We used an fMRI data set of 24 participants imagining possible future events that had been categorized *a priori* as high versus low in concreteness and positive versus negative in valence (Lee et al., 2020, **Fig 3-2A**). We used thresholded partial least squares (TPLS, see **Fig. 3-1**) to develop a whole-brain classifier that discriminated events that were high versus low concreteness. We checked by cross-validation within the training dataset to ensure that our predictor of concreteness could accurately, out-of-sample, predict the concreteness but not the valence of imagined future events. As expected, the neural predictor of imagination concreteness successfully discriminated the trial categories of high versus low concreteness (mean prediction AUC = 62.56%, t-test against 50%, $t(23) = 7.38$, $p < .001$), but not the trial categories of positive versus negative valence (mean prediction AUC = 51.12%, t-test against 50%, $t(23) = 0.73$, $p = .47$; **Fig. 3-2B**). We also further checked whether our predictions were also aligned with the participants' ratings of the concreteness of imagined future events but not with their ratings of valence. Again, we found that our predictor was able to predict out-of-sample ratings of concreteness but unable to predict ratings of valence (**Fig. 3-2C**). Mean out-of-sample correlation between the neural prediction score and concreteness ratings was $r = 0.15$ ($t(23) = 4.80$, $p < .001$), while the correlation between the neural prediction score and valence ratings was $r = -0.0069$ ($t(23) = -0.26$, $p = 0.79$).

The whole-brain prediction map of concreteness involved various regions of the brain, mostly in a bilateral fashion (**Fig. 3-3 & Table 3-1**). Positive coefficients (predictive of higher concreteness) were found in bilateral hippocampus, bilateral central orbitofrontal cortex (OFC), posterior cingulate cortex (PCC), bilateral middle occipital gyri, left dorsolateral prefrontal cortex (dlPFC), and medial posterior OFC. Negative coefficients (predictive of lower concreteness) were

found in bilateral temporal poles, dorsomedial prefrontal cortex (dmPFC), bilateral temporoparietal junction (TPJ), bilateral middle temporal gyri, and precuneus.

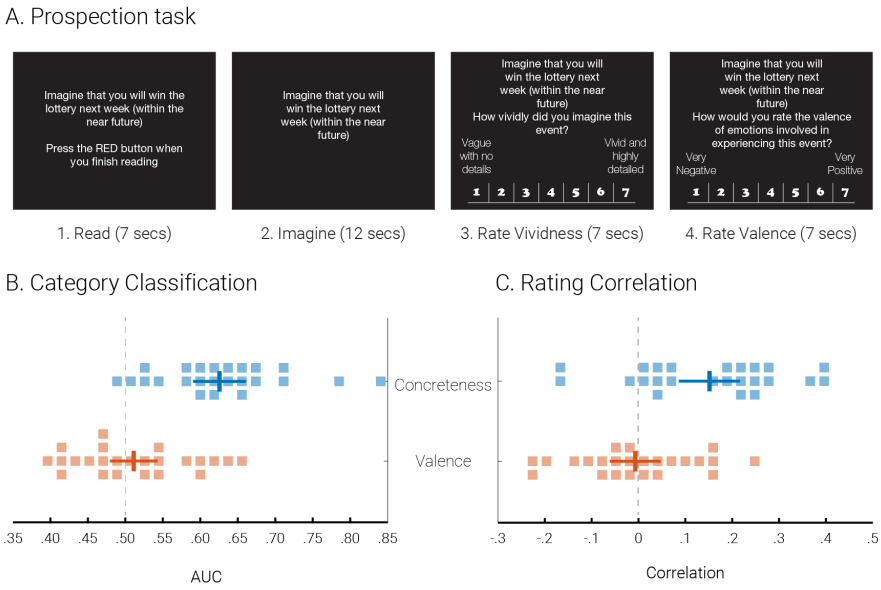


**Figure 3 - 2. Out-of-sample prediction of concreteness and valence in prospection dataset achieved by 24-fold leave-one-out cross validation.**

Panel A shows the schematic of the prospection task from Lee et al. (2020). A whole-brain concreteness predictor is trained on 23 people's data and used to predict the left-out person's data. Panel B shows classification performance on a priori trial categories of concreteness (high versus low) and valence (positive versus negative) as measured by area under the receiver operating characteristic curve. Panel C shows correlation with concreteness and valence ratings provided by participants. Each dot represents one participant, the vertical bar represents the mean, and the horizontal bar represents the 95% confidence interval of the mean.

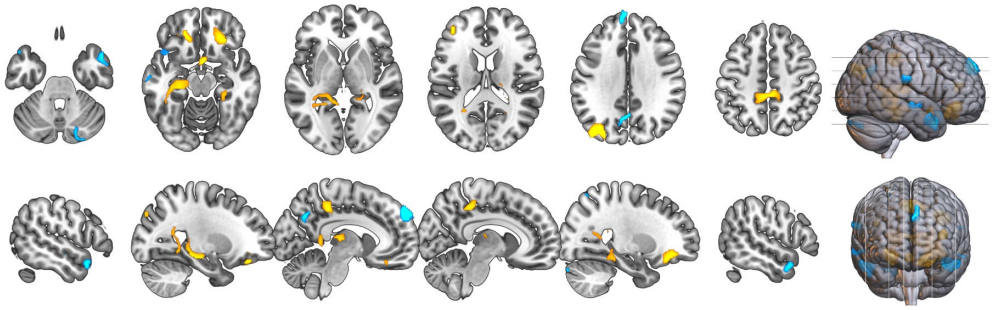

**Figure 3 - 3. Whole-brain predictor of the concreteness of imagined future events built from 24 participants.**

The warm colors indicate positive coefficients and cool colors indicate negative coefficients. Notable regions include bilateral central OFC, PCC, and bilateral hippocampus for positive coefficients, and dmPFC, precuneus, and bilateral temporal pole for negative coefficients.

| Description | Size (number of voxels) | X | Y | Z |
|---|---|---|---|---|
| *Positive* | | | | |
| Bilateral Hippocampus | 1124 voxels (left & right) | -27 | -30 | -12 |
| Bilateral Central Orbitofrontal Cortex | 169 voxels (left) | -25 | 35 | -18 |
| | 306 voxels (right) | 24 | 35 | -12 |
| Posterior Cingulate Cortex | 379 voxels (medial) | 4 | -36 | 45 |
| Bilateral Middle Occipital Gyri | 302 voxels (left) | -37 | -82 | 33 |
| | 15 voxels (right) | 44 | -72 | 23 |
| Left Dorsolateral Prefrontal Cortex (left middle frontal gyrus) | 70 voxels (left) | -39 | 37 | 11 |
| Medial Posterior Orbitofrontal Cortex | 57 voxels (medial) | 2 | 3 | -18 |
| *Negative* | | | | |
| Bilateral Temporal Pole | 207 voxels (left) | -45 | 19 | -26 |
| | 184 voxels (right) | 54 | 9 | -26 |
| Dorsomedial Prefrontal Cortex (medial superior frontal gyri) | 104 voxels (medial) | -5 | 59 | 39 |
| Bilateral Temporoparietal Junction | 11 voxels (left) | -61 | -24 | 21 |
| | 82 voxels (right) | 66 | -24 | 23 |
| Bilateral Middle Temporal Gyri | 25 voxels (left) | -61 | -14 | -14 |
| | 67 voxels (right) | 62 | -10 | -8 |
| Precuneus | 46 voxels (medial) | 2 | -58 | 33 |

**Table 3 - 1. Clusters of whole-brain predictor of imagination concreteness.**

Clusters of voxels that have non-zero coefficients in the final predictor of the concreteness of imagined future events are reported, grouped by the sign of the coefficients and ordered by cluster size. From left to right, the region names, cluster size in voxels, and MNI coordinates are provided. Clusters that are 3 voxels or smaller are not reported.

We next applied this whole-brain predictor of concreteness in two separate delay discounting tasks, in order to test whether the neural signature of concrete future thinking was higher when considering sooner rewards and lower when considering later rewards. In both datasets we found that neural concreteness scores were negatively correlated with the delay until the receipt of the reward, such that farther delays were associated with lower concreteness scores. Firstly, in an intertemporal bidding task, participants (n = 39) were presented with a fixed monetary outcome of $75 at different delays and asked to report the immediate amount they

would feel to be equivalent to the delayed outcome. For each trial, we calculated neural concreteness scores by applying the whole brain predictor developed above to the activity for that trial. We found that the trial-by-trial neural concreteness scores were correlated negatively with delay (**Fig. 3-4**; mean $r$ = -0.060, $t(38)$ = -3.79, $p$ < .001), such that shorter delays (i.e., more proximal future) were associated with higher concreteness scores, and longer delays (i.e., more distant future) with lower concreteness scores.

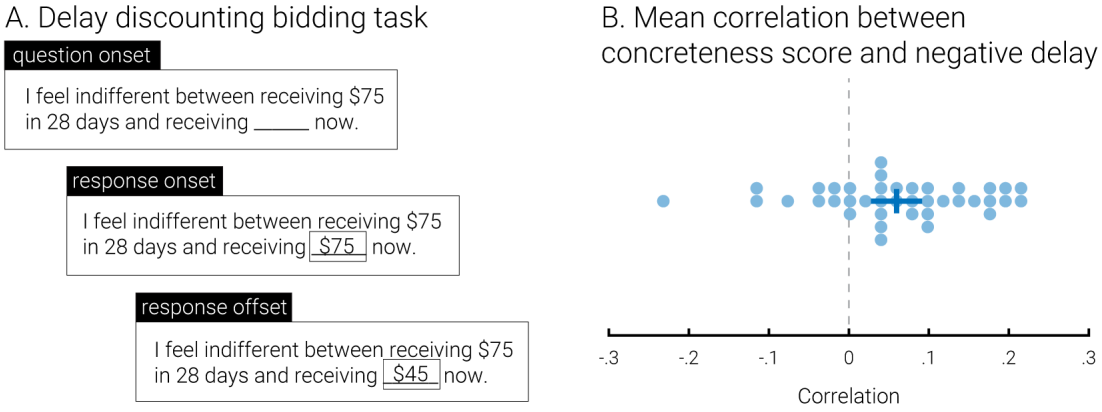


**Figure 3 - 4. Out-of-sample prediction of delay in an intertemporal bidding task.**

Panel A shows the bidding task structure from Cooper et al. (2013). Participants are first shown the delayed amount of $75 (fixed) and a variable delay and are asked to bid their immediate equivalent. Panel B shows the per-person correlation between trial-by-trial delay (sign-flipped) and concreteness prediction scores from the whole-brain predictor. The vertical bar represents the mean, and the horizontal bar represents the 95% confidence interval of the mean (n = 39).

We replicated this finding in a second dataset in which participants made discrete binary choices between immediate and delayed rewards. In this choice task from Kable et al. (2017), participants (n = 166) made choices between a fixed immediate reward of $20 and a future reward that varied in amount ($21 ~ $85) and delay (20 ~ 180 days) across trials. Again, we found that the trial-by-trial neural concreteness scores were correlated negatively with delay (**Fig. 3-5**; mean $r$ = -0.070, $t(165)$ = -9.23, $p$ < .001), such that shorter delays were associated with higher

concreteness scores. Furthermore, this association was specific to the delay to reward. Although the neural concreteness scores were also positively correlated with the delayed amount (mean $r$ = 0.037, $t(165) = 4.25$, $p < .001$), concreteness was more strongly associated with delay than amount (paired $t$-test, $t(165) = 3.15$, $p = 0.0019$).

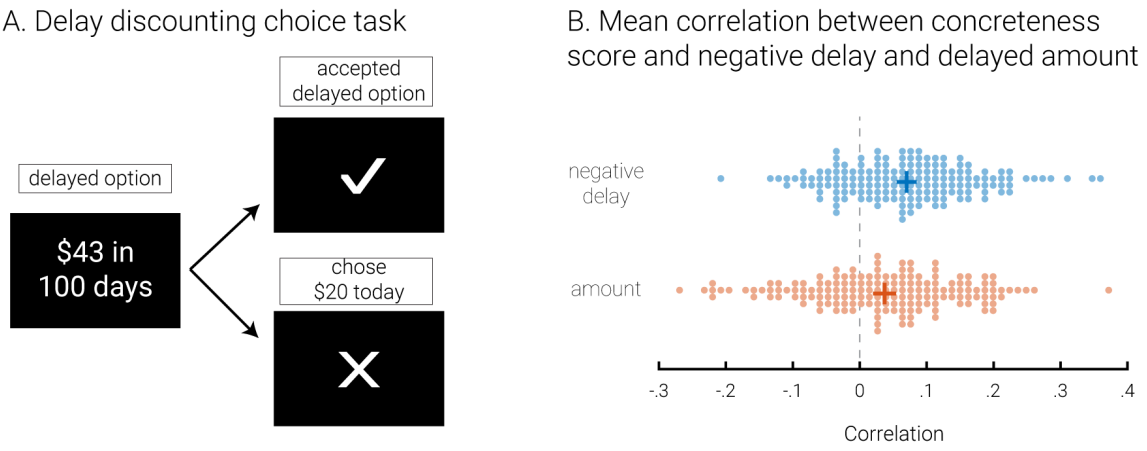


**Figure 3 - 5. Out-of-sample prediction of delay in an intertemporal choice task.**

Panel A shows the choice task structure from Kable et al. (2017). Participants are shown the delayed reward and are asked to either accept it or to reject it for $20 immediately. Panel B shows the per-person correlation between trial-by-trial delay (sign-flipped) and concreteness prediction scores in comparison to that between trial-by-trial amount and concreteness prediction scores (delay has been sign-flipped to facilitate this comparison). The vertical bar represents the mean, and the horizontal bar represents the 95% confidence interval of the mean (n = 166).

## Discussion

Multiple theories in psychology have suggested that delayed outcomes are discounted in value relative to immediate outcomes in part because more temporally distant options are perceived as less concrete and tangible than more temporally proximal ones (Liberman & Trope, 2014; Rick & Loewenstein, 2008; Trope & Liberman, 2010). These theories have been supported by a range of via various behavioral experiments (Bischoff & Hansen, 2016; Kelley & Schmeichel, 2015; Liberman & Trope, 2014; Malkoc, Zauberman, & Bettman, 2010; Mischel &

Baker, 1975; Yi, Stuppy-Sullivan, Pickover, & Landes, 2017). Here we add converging neuroscientific evidence to these theories. We used fMRI data during an imagination task to create a whole-brain, multivariate predictor specific to the concreteness of prospective thought, independent of the valence of prospective thought. Then we show, in two separate delay discounting datasets with markedly different task structure (one bidding task, one choice task), that the neural signature of concreteness is modulated by the temporal distance to the delayed option under consideration. That is, the pattern of neural activity that predicts more concrete prospective thinking is stronger for more temporally proximal outcomes and weaker for more temporally distal ones. The neural signature of concreteness was also more strongly modulated by the delay to reward than by the magnitude of reward. These results show that, while people are making intertemporal decisions, an online, unobtrusive neural index of concrete thinking declines as the outcomes considered are delayed farther into the future.

Our results complement previous tests of construal level theory using fMRI. These studies have shown that neural activity associated with imagining near events, compared to distant events, overlaps with neural activity engaged by other forms of psychological proximity or by low- versus high-level construal (Stillman et al., 2017; Tamir & Mitchell, 2011). Here we make two advances over these previous results. First, we distinguish between neural activity due to the concreteness, versus the valence, of prospective thought. This is critical as several previous studies have found the strongest increases in activity for sooner, compared to later, events in the mPFC and PCC (Mitchell et al., 2011; Tamir & Mitchell, 2011), two regions that we have previously shown are associated with the valence of prospective thought (Lee et al., 2020). Second, we show that a neural index of concreteness is modulated by the delay to the outcome *during intertemporal decision-making*. This links reduced concreteness directly to the discounting of future rewards, a process known to be associated with many important life outcomes (Alessi &

Petry, 2003; Anderson & Mellor, 2008; Brañas-Garza et al., 2007; Kirby et al., 1999; Krain et al., 2008; Lejuez et al., 2005, 2003; Schepis et al., 2011; Shamosh & Gray, 2008).

The whole-brain prediction map for concreteness is remarkably consistent with findings from other lines of research. Several previous studies have argued that the orbitofrontal represents the features of potential outcomes during decision making (Burke, Franz, Miller, & Schoenbaum, 2008; Howard, Gottfried, Tobler, & Kahnt, 2015; Y. K. Takahashi et al., 2013), and that interactions with the hippocampus may be critical for generating these representations from memory (for review, see Shohamy & Daw, 2015). Furthermore, there is evidence that these regions play a role in valuing delayed rewards. Lesions to the OFC caused increased impatience (Sellitto, Ciaramelli, & Di Pellegrino, 2010), and reduced grey matter thickness in both the OFC (Pehlivanova et al., 2018) and the medial temporal lobe (Lempert et al., 2020; Owens et al., 2017) is associated with increased discounting. Correspondingly, we would expect that modulating activity in these regions as people consider future outcomes would alter the concreteness with which those outcomes are imagined and the degree to which those outcomes are discounted.

To obtain the current results, we applied a novel adaptation of partial least squares optimized to construct interpretable whole-brain predictors with minimal computation time (study 1). Though many different methods for constructing whole brain predictors have been proposed (Grosenick et al., 2008; Kragel & LaBar, 2014; Smith et al., 2014; Wager et al., 2013), none have yet achieved widespread use in the field. Here we illustrate what we think is the most promising and exciting potential use of such predictors: decoding mental states online in order to test psychological hypotheses.

**Supplemental Materials**

*Image preprocessing for prospection dataset*

Results included in this manuscript come from preprocessing performed using fMRIPrep 20.0.7 (Esteban, Markiewicz, et al. (2018); Esteban, Blair, et al. (2018); RRID:SCR_016216), which is based on Nipype 1.4.2 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

*Anatomical data preprocessing.* The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al. 2010), distributed with ANTs 2.2.0 (Avants et al. 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang, Brady, and Smith 2001). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym],

*Functional data preprocessing*. For each of the 4 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-

stripped version were generated using a custom methodology of fMRIPrep. A B0-nonuniformity map (or fieldmap) was estimated based on a phase-difference map calculated with a dual-echo GRE (gradient-recall echo) sequence, processed with a custom workflow of SDCFlows inspired by the epidewarp.fsl script and further improvements in HCP Pipelines (Glasser et al. 2013). The fieldmap was then co-registered to the target EPI (echo-planar imaging) reference run and converted to a displacements field map (amenable to registration tools such as ANTs) with FSL's fugue and other SDCflows tools. Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using flirt (FSL 5.0.9, Jenkinson and Smith 2001) with the boundary-based registration (Greve and Fischl 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox and Hyde 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al. 2014). The three

global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al. 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al. 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

*Image preprocessing for delay discounting bidding dataset*

Results included in this manuscript come from preprocessing performed using fMRIPrep 20.1.1 (Esteban, Markiewicz, et al. (2018); Esteban, Blair, et al. (2018); RRID:SCR_016216), which is based on Nipype 1.5.0 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

*Anatomical data preprocessing.* A total of 1 T1-weighted (T1w) images were found within the input BIDS dataset.The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al. 2010), distributed with ANTs 2.2.0 (Avants et al. 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang, Brady, and Smith 2001). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym],

*Functional data preprocessing.* For each of the 4 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox and Hyde 1997, RRID:SCR_005927). Susceptibility distortion correction (SDC) was omitted. The BOLD reference was then co-registered to the T1w reference using flirt (FSL 5.0.9, Jenkinson and Smith 2001) with the boundary-based registration (Greve and Fischl 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al. (2014)) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al. (2002)). FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al. 2007). Principal components are estimated after high-pass filtering the

preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al. 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

*Image preprocessing for delay discounting choice dataset*

Results included in this manuscript come from preprocessing performed using fMRIPrep 20.0.5 (Esteban, Markiewicz, et al. (2018); Esteban, Blair, et al. (2018); RRID:SCR_016216), which is based on Nipype 1.4.2 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

*Anatomical data preprocessing.* A total of 2 T1-weighted (T1w) images were found within the input BIDS dataset. All of them were corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al. 2010), distributed with ANTs 2.2.0 (Avants et al. 2008, RRID:SCR_004757). The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang, Brady, and Smith 2001). A T1w-reference map was computed after registration of 2 T1w images (after INU-correction) using mri_robust_template (FreeSurfer 6.0.1, Reuter, Rosas, and Fischl 2010). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym],

*Functional data preprocessing.* For each of the 18 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-

stripped version were generated using a custom methodology of fMRIPrep. A B0-nonuniformity map (or fieldmap) was estimated based on a phase-difference map calculated with a dual-echo GRE (gradient-recall echo) sequence, processed with a custom workflow of SDCFlows inspired by the epidewarp.fsl script and further improvements in HCP Pipelines (Glasser et al. 2013). The fieldmap was then co-registered to the target EPI (echo-planar imaging) reference run and converted to a displacements field map (amenable to registration tools such as ANTs) with FSL's fugue and other SDCflows tools. Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using flirt (FSL 5.0.9, Jenkinson and Smith 2001) with the boundary-based registration (Greve and Fischl 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox and Hyde 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al. 2014). The three

global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al. 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al. 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

# CHAPTER 5 - GENERAL DISCUSSION

This dissertation examined a novel construction algorithm of whole-brain decoders and how whole-brain decoders can aid neuroscientific and psychological research. The overarching goal of the three studies has been to provide an easy to implement method for whole-brain decoder construction and to demonstrate its potential uses in substantive research. The first study addressed two common difficulties with creating a whole-brain decoder: interpretability and computational burden. By exploiting analytical properties of partial least squares, I proposed Thresholded Partial Least Squares (TPLS), which can create interpretable whole-brain decoders in much shorter time compared to existing approaches and also yield interpretable decoders. In coming years, as neuroimaging datasets grow larger in size, I hope that provided algorithm and the statistical package will be useful for many researchers.

In the second study, I demonstrated how an interpretable whole-brain decoder can be interpreted. I built a whole-brain decoder of valuation based on two large behavioral economic decision-making tasks. Thanks to the interpretability of the whole-brain decoder, I found that numerous regions were shown to be negatively predictive of value-based choice despite the fact that no regions were negatively correlated with said choice. This provided the hint that there may be common 'noise' among the value signals that the whole-brain decoder attempted to cancel out by pitting regions against each other with opposing signs. When I decomposed valuation signals in the brain into signals related to valuation of the current trial versus past trials, I found that neural signals of valuation in the brain had history dependency while behavior did not. More specifically, neural signals of value varied in magnitude depending on the valuation of the past trials' options, while participants' choices regarding the current trial did not depend on the valuation of the past trials' options. While this does not necessarily implicate that our brain

actually renders decisions based on combinations of multiple regions' signal, it does showcase how whole-brain decoders combine signals across the brain and how examining it can yield novel insights about the valuation signal.

In the third study, I show an example of how whole-brain decoders can empirically measure mental states and processes that are difficult to probe otherwise. Several psychological theories have posited that one possible reason for discounting delayed rewards is because future rewards are imagined less vividly than immediate rewards (Liberman & Trope, 2014; Rick & Loewenstein, 2008; Trope & Liberman, 2010). I provide neural evidence for this claim by constructing a whole-brain decoder of imagination vividness and showing that indeed, when the decoder is applied to peoples' brain images in delay discounting tasks, peoples' imagination vividness scores are negatively correlated with the delay until the receipt of the reward. This study, in its simple design, demonstrates the benefit of generalized brain-decoders that can empirically measure and predict private mental processes that are otherwise hard to measure.

While the dissertation has highlighted the substantive benefits gained from the whole-brain decoders, there can be many practical benefits to be gained by the whole-brain decoder of valuation and imagination vividness, which were constructed in this dissertation. In business research, accurate forecasting of the market's preference for new items is critically important, and recent research has shown promising signs that the market-level preference can be predicted from neural readings of smaller test groups (Berns & Moore, 2012; Genevsky et al., 2017; Venkatraman et al., 2014). While many of these studies have used the partial-brain approach of prediction, a generalized whole-brain decoder of value could prove to be more robust in predicting market-level preferences. Of course, it is possible that the partial-brain predictor of market-level preference may be better than the whole-brain predictor of individual preferences, at least for market-level prediction. In such cases, one could consider directly constructing the

whole-brain decoder of market-level preference and comparing it against that of individual preference to see how they differ. Outside business research, the predictors can also be useful in communications research where mental responses elicited from public messages or advertisements can be predictive of subsequent behavioral change. For example, in a recent study, Schmälzle et al. (2020) have found that anti-smoking web banners that measured as being more negatively valenced and more vivid, based on the predictors created from study 3, were less likely to be clicked at the population level, suggesting that whole-brain decoder of mental processes can provide valuable tools of predicting population-level behavior.

It is important to note that the whole-brain decoding method, presented here, is not meant to be a substitute to traditional mass-univariate analysis; rather, it should be used in conjunction to yield insight. Traditional fMRI analysis methods, using mass-univariate analysis, assess the correlation between the signal of interest and voxel activity. The whole-brain decoder presented here, attempts to predict the signal of interest using combinations of voxel activity. As in the case with regression, predictive voxels that are positively correlated with signal of interest may have negative coefficients when entered into a regression, depending on other variables. This difference is most well exemplified in study 2, where several regions were given negative coefficients in the whole-brain predictor of value-based choice, while traditional mass-univariate analysis led to positive coefficients for all regions. This difference was the key insight in understanding that 1) all value signals are similar to each other, but 2) pitting them against each other may help prediction by canceling out some noise. Hence, while the interpretability of the whole-brain decoder allows for identification of predictive regions and their regression coefficients, the predictor must be understood and interpreted together with traditional univariate methods as well.

Additionally, the whole-brain decoding method does not supersede partial-brain prediction methods either. Again, they should be used in conjunction with each other. Partial-brain decoders (i.e., typical MVPA methods) excel in localizing the signal and answering a priori questions about whether a given region can predict/decode certain mental states or behavior. Whole-brain decoders, on the other hand, only deal with whether and how the entire brain can predict a given mental state or behavior. Hence, it is feasible that regions identified as being predictive in a whole-brain decoder may not be able to predict mental states or behavior by itself. This may be either due to the signal within the region not being strong enough to be standalone predictors, or it may be due to that region only being useful in prediction in combination with other regions.

One important caveat of the whole-brain decoder approach is that, in contrast to mass-univariate methods, it does not provide measures of variability or statistical significance tests. This is not specific to whole-brain decoders but is true for most modern regression algorithms: while the models are fit using cross-validation so as to maximize out-of-sample predictive performance, there's no guarantee about the stability of the fitted models as they are subject to change should one choose a different cross-validation scheme. Given the lack of consensus on what a best cross-validation scheme is, especially in neuroimaging, a larger dataset would be better at dealing with this problem as it would be less subject to spurious changes in signal due to cross-validation data splits. Regardless, care should be taken in interpreting the whole-brain decoders as they are not a result of a statistical test, unlike typical mass-univariate analysis maps. Furthermore, it can be useful to consider sensitivity analysis using various cross-validation schemes to check the robustness of the results.

In future research, correlational relationship between different whole-brain decoders can be investigated to serve as a measure of 'neural distance', which reflect how much two different

mental processes share similar brain activities. Whole-brain decoder of different tasks may be used to measure neural similarities between tasks and validated against behavioral similarities to examine how different tasks are related to one another. Such research may shed insight into why certain mental disorders are likely to impact certain sets of behaviors together more so than other behaviors. In alternate use of whole-brain decoders, they can also be constructed in two different sub-populations (e.g., healthy and clinical) to compare the mental processes between them. In clinical research, such characterization of abnormal mental processes via neural signatures can prove to be a useful diagnostic tool that can detect psychological abnormalities in neural functions even before behavioral manifestations. If individually-tailored whole-brain decoders are possible, future research may apply clustering algorithms across peoples' brain decoders to identify subgroups and heterogeneity of mental processes within the same task.

Whole-brain decoders, as sci-fi as it sounds, have a lot of potential use in neuroscientific and psychological research, both substantive and applied. Not only do we get to learn about the brain in ways we did not before, we also get the closest thing to mind reading. After all, haven't we all thought to ourselves, 'if only I could read minds?' I hope this dissertation invites and convinces more researchers to consider using whole-brain decoders in their research and someday we might actually get to read minds like a book.

BIBLIOGRAPHY

Alessi, S. M., & Petry, N. M. (2003). Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behavioural Processes*, *64*(3), 345–354.

Anderson, L. R., & Mellor, J. M. (2008). Predicting health behaviors with an experimental measure of risk preference. *Journal of Health Economics*, *27*(5), 1260–1274.

Barron, H. C., Dolan, R. J., & Behrens, T. E. J. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience*. https://doi.org/10.1038/nn.3515

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427.

Berns, G. S., & Moore, S. E. (2012). A neural predictor of cultural popularity. *Journal of Consumer Psychology*, *22*(1), 154–160. https://doi.org/10.1016/j.jcps.2011.05.001

Bischoff, C., & Hansen, J. (2016). Influencing support of charitable objectives in the near and distant future: delay discounting and the moderating influence of construal level. *Social Influence*. https://doi.org/10.1080/15534510.2016.1232204

Brañas-Garza, P., Georgantzís, N., & Guillén, P. (2007). Direct and indirect effects of pathological gambling on risk attitudes. *Judgment and Decision Making*, *2*(2), 126–136.

Breiter, H. C., Gollub, R. L., Weisskoff, R. M., Kennedy, D. N., Makris, N., Berke, J. D., … Hyman, S. E. (1997). Acute effects of cocaine on human brain activity and emotion. *Neuron*, *19*(3), 591–611. https://doi.org/10.1016/S0896-6273(00)80374-8

Burke, K. A., Franz, T. M., Miller, D. N., & Schoenbaum, G. (2008). The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. *Nature*. https://doi.org/10.1038/nature06993

Cai, X., & Padoa-Schioppa, C. (2012). Neuronal Encoding of Subjective Value in Dorsal and Ventral Anterior Cingulate Cortex. *Journal of Neuroscience*. https://doi.org/10.1523/jneurosci.3864-11.2012

Carter, R. M. K., Meyer, J. R., & Huettel, S. A. (2010). Functional Neuroimaging of Intertemporal Choice Models: A Review. *Journal of Neuroscience, Psychology, and Economics*. https://doi.org/10.1037/a0018046

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., & Wager, T. D. (2015). A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biology*, *13*(6), 1–28. https://doi.org/10.1371/journal.pbio.1002180

Cloutier, J., Heatherton, T. F., Whalen, P. J., & Kelley, W. M. (2008). Are attractive people rewarding? Sex differences in the neural substrates of facial attractiveness. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn.2008.20062

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*. https://doi.org/10.1126/science.aaf0941

Cooper, N., Kable, J. W., Kim, B. K., & Zauberman, G. (2013). Brain activity in valuation regions while thinking about the future predicts individual discount rates. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0400-13.2013

Cox, K. M., & Kable, J. W. (2014). BOLD Subjective Value Signals Exhibit Robust Range Adaptation. *The Journal of Neuroscience*, *34*(49), 16533–16543.

https://doi.org/10.1523/JNEUROSCI.3927-14.2014

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*. https://doi.org/10.1016/0169-7439(93)85002-X

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1113427109

DeWitt, I., & Rauschecker, J. P. (2013). Wernicke's area revisited: Parallel streams and word processing. *Brain and Language*. https://doi.org/10.1016/j.bandl.2013.09.014

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*. https://doi.org/10.1038/nature08704

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*. https://doi.org/10.1038/s41592-018-0235-4

Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2013.03.041

Falk, E. B., Berkman, E. T., & Lieberman, M. D. (2012). From Neural Responses to Population Behavior: Neural Focus Group Predicts Population-Level Media Effects. *Psychological Science*. https://doi.org/10.1177/0956797611434964

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*.

https://doi.org/10.18637/jss.v033.i01

Genevsky, A., & Knutson, B. (2015). Neural affective mechanisms predict market-level microlending. *Psychological Science*. https://doi.org/10.1177/0956797615588467

Genevsky, A., Yoon, C., & Knutson, B. (2017). When brain beats behavior: Neuroforecasting crowdfunding outcomes. *The Journal of Neuroscience*, 1633–16. https://doi.org/10.1523/JNEUROSCI.1633-16.2017

Goldstein, W. M., & Einhorn, H. J. (1987). Expression Theory and the Preference Reversal Phenomena. *Psychological Review*, *94*(2), 236–254.

Green, L., Fry, A. F., & Myerson, J. (1994). Research Report Discounting of Delayed Rewards:A Life-Span Comparison. *Pschological Science*, *5*(1), 33–36.

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.130.5.769

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2005.11.006

Grosenick, L., Greer, S., & Knutson, B. (2008). Interpretable classi ers for FMRI improve prediction of purchases. *Analysis*, *X*(Xx), 1–10.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., & Taylor, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, *72*, 304–321. https://doi.org/10.1016/j.neuroimage.2012.12.062

Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding Structure with Randomness :

Probabilistic Algorithms for Matrix Decompositions. *SIAM Review*.

Herrnstein, R. J. (1981). Self-control as response strength. *Quantification of Steady-State Operant Behavior*, 3–20.

Howard, J. D., Gottfried, J. A., Tobler, P. N., & Kahnt, T. (2015). Identity-specific coding of future rewards in the human orbitofrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1503550112

Izuma, K., Saito, D. N., & Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive Neuroscience*, *22*(4), 621–631. https://doi.org/10.1162/jocn.2009.21228

Jung, W. H., Lee, S., Lerman, C., & Kable, J. W. (2018). Amygdala Functional and Structural Connectivity Predicts Individual Risk Tolerance. *Neuron*, *98*(2), 394-404.e4. https://doi.org/10.1016/j.neuron.2018.03.019

Kable, J. W., Caulfield, M. K., Falcone, M., McConnell, M., Bernardo, L., Parthasarathi, T., … Hornik, R. (2017). No effect of commercial cognitive training on neural activity during decision-making. *Journal of Neuroscience*, 2816–2832.

Kable, J. W., Caulfield, M. K., Falcone, M., McConnell, M., Bernardo, L., Parthasarathi, T., … Lerman, C. (2017). No Effect of Commercial Cognitive Training on Brain Activity, Choice Behavior, or Cognitive Performance. *The Journal of Neuroscience*, *37*(31), 7390–7402.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value druing intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633.

Kahneman, D., & Tversky, A. (1979). Kahneman & Tversky (1979) - Prospect Theory - An

Analysis Of Decision Under Risk. *Econometrica*. https://doi.org/10.2307/1914185

Kanske, P., Böckler, A., Trautwein, F. M., Lesemann, F. H. P., & Singer, T. (2016). Are strong empathizers better mentalizers? Evidence for independence and interaction between the routes of social cognition. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsw052

Kanske, P., Böckler, A., Trautwein, F. M., & Singer, T. (2015). Dissecting the social brain: Introducing the EmpaToM to reveal distinct neural networks and brain-behavior relations for empathy and Theory of Mind. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2015.07.082

Karmarkar, U. R., Shiv, B., & Knutson, B. (2015). Cost Conscious? The Neural and Behavioral Impact of Price Primacy on Decision Making. *Journal of Marketing Research*, *52*(4), 467–481. https://doi.org/10.1509/jmr.13.0488

Kelley, N. J., & Schmeichel, B. J. (2015). Thinking about Death Reduces Delay Discounting. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0144228

Khaw, M. W., Glimcher, P. W., & Louie, K. (2017). Normalized value coding explains dynamic adaptation in the human valuation process. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1715293114

Kirby, K. N., Petry, Nancy, M., & Bickel, Warren, K. (1999). Heroin addicts have higher discount rates for delayed rewards than non drug using controls. *Journal of Experimental Psychology*, *128*(1), 78–87.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, *21*(16),

RC159–RC159.

Knutson, B., Scott, R., Wimmer, G. E., Prelec, D., & Loewenstein, G. (2007). Neural Predictors of Energy-Efficient Purchases. *Neuron*, *53*(1), 147–156.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *Journal of Neuroscience*, *25*(19), 4806–4812.

Kobayashi, S., Pinto de Carvalho, O., & Schultz, W. (2010). Adaptation of Reward Sensitivity in Orbitofrontal Neurons. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.4009-09.2010

Kragel, P. A., & LaBar, K. S. (2014). Multivariate neural biomarkers of emotional states are categorically distinct. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsv032

Krain, A. L., Gotimer, K., Hefton, S., Ernst, M., Castellanos, F. X., Pine, D. S., & Milham, M. P. (2008). A Functional Magnetic Resonance Imaging Investigation of Uncertainty in Adolescents with Anxiety Disorders. *Biological Psychiatry*, *63*(6), 563–568.

Kringelbach, M. L., O'Doherty, J., Rolls, E. T., & Andrews, C. (2003). Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness. *Cerebral Cortex*, *13*(10), 1064–1071.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, *112*(2), 443–478.

Lee, S., Glaze, C. M., Bradlow, E. T., & Kable, J. (2019). Agnostic Utility Function Approximation via Cubic Bezier Splines. https://doi.org/10.31234/osf.io/bv2gk

Lee, S., Lerman, C., & Kable, J. W. (2019). Neural Correlates of Value Are Intrinsically History Dependent. *NEURON-D-19-01311*.

Lee, S., Parthasarathi, T., & Kable, J. W. (2020). The dorsal and ventral default mode networks are dissociably modulated by the valence and vividness of imagined events. *BioRxiv*.

Lejuez, C. W., Aklin, W. M., Bornovalova, M. A., & Moolchan, E. T. (2005). Differences in risk-taking propensity across inner- city adolescent ever-and never-smokers. *Nicotine & Tobacco Reserach*, *7*(1), 71–79.

Lejuez, C. W., Aklin, W. M., Jones, H. A., Strong, D. R., Richards, J. B., Kahler, C. W., & Read, J. P. (2003). The Balloon Analogue Risk Task (BART) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology*, *11*(1), 26–33.

Lempert, K. M., Mechanic-Hamilton, D. J., Xie, L., Wisse, L. E. M., de Flores, R., Wang, J., … Kable, J. W. (2020). Neural and behavioral correlates of episodic memory are associated with temporal discounting in older adults. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2020.107549

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of Significance: Principal component analysis. *Nature Methods*. https://doi.org/10.1038/nmeth.4346

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038.

Levy, I., Lazzaro, S. C., Rutledge, R. B., & Glimcher, P. W. (2011). Choice from non-choice: predicting consumer preferences from blood oxygenation level-dependent signals obtained during passive viewing. *Journal of Neuroscience*, *31*(1), 118–125.

Liberman, N., & Trope, Y. (2014). Traversing psychological distance. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2014.03.001

Malkoc, S. A., Zauberman, G., & Bettman, J. R. (2010). Unstuck from the concrete: Carryover effects of abstract mindsets in intertemporal preferences. *Organizational Behavior and Human Decision Processes*. https://doi.org/10.1016/j.obhdp.2010.07.003

Markowitz, H. (1959). Portfolio selection: efficient diversification of investments. New Haven, CT: Cowles Foundation.

McLean, J., Brennan, D., Wyper, D., Condon, B., Hadley, D., & Cavanagh, J. (2009). Localisation of regions of intense pleasure response evoked by soccer goals. *Psychiatry Research - Neuroimaging*, *171*(1), 33–43. https://doi.org/10.1016/j.pscychresns.2008.02.005

Mischel, W., & Baker, N. (1975). Cognitive appraisals and transformations in delay behavior. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/h0076272

Mitchell, J. P., Schirmer, J., Ames, D. L., & Gilbert, D. T. (2011). Medial prefrontal cortex predicts intertemporal choice. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn.2010.21479

Morgenstern, O., & Von Neumann, J. (1953). *Theory of games and economic behavior*. Princeton university press.

Naccache, L., & Dehaene, S. (2001). The priming method: imaging unconscious repetition priming reveals an abstract …. *Cerebral Cortex*.

O'Doherty  Deichmann, R, Critchley, HD, Dolan, RJ, J. P. (2002). Neural responses during antcipation of a primary taste reward. *Neuron*, *33*(5), 815–826.

Owens, M. M., Gray, J. C., Amlung, M. T., Oshri, A., Sweet, L. H., & MacKillop, J. (2017). Neuroanatomical foundations of delayed reward discounting decision making. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.08.045

Padoa-Schioppa, C. (2009). Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.3751-09.2009

Pegors, T. K., Kable, J. W., Chatterjee, A., & Epstein, R. A. (2015). Common and unique representations in pFC for face and place attractiveness. *Journal of Cognitive Neuroscience*, *27*(5), 959–973. https://doi.org/10.1162/jocn_a_00777

Pehlivanova, M., Wolf, D. H., Sotiras, A., Kaczkurkin, A. N., Moore, T. M., Ciric, R., … Satterthwaite, T. D. (2018). Diminished cortical thickness is associated with impulsive choice in adolescence. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2200-17.2018

Plassmann, H., O'Doherty, J. P., & Rangel, A. (2010). Appetitive and Aversive Goal Values Are Encoded in the Medial Orbitofrontal Cortex at the Time of Decision Making. *Journal of Neuroscience*, *30*(32), 10799–10808. https://doi.org/10.1523/JNEUROSCI.0788-10.2010

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. *Journal of Neuroscience*, *27*(37), 9984–9988. https://doi.org/10.1523/JNEUROSCI.2131-07.2007

Polanía, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*. https://doi.org/10.1038/s41593-018-0292-0

Prelec, D. (1998). The Probability Weighting Function. *Econometrica*, *66*(3), 497.

Price, C. J. (2012). A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2012.04.062

Qian, J., Hastie, T., Friedman, J., Tibshirani, R., & Simon, N. (2013). Glmnet for Matlab, 2013. *URL Http://Www. Stanford. Edu/~ Hastie/Glmnet_matlab*.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Rick, S., & Loewenstein, G. (2008). Review. Intangibility in intertemporal choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2008.0150

Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2004.06.035

Samuelson, P. a. (1937). Note on Measurement of Utility. *The Review of Economic Studies*, *4*(2), 155–161.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*. https://doi.org/10.1016/S1053-8119(03)00230-1

Saxe, Rebecca, & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Schepis, T. S., McFetridge, A., Chaplin, T. M., Sinha, R., & Krishnan-Sarin, S. (2011). A pilot examination of stress-related changes in impulsivity and risk taking as related to smoking status and cessation outcome in adolescents. *Nicotine and Tobacco Research*, *13*(7), 611–615.

Schmälzle, R., Cooper, N., O'Donnell, M. B., Tompson, S., Lee, S., Cantrell, J., … Falk, E. B. (2020). The Effectiveness of Online Messages for Promoting Smoking Cessation Resources: Predicting Nationwide Campaign Effects From Neural Responses in the EX Campaign. *Frontiers in Human Neuroscience*, *14*. https://doi.org/10.3389/fnhum.2020.565772

Scholz, C., Baek, E. C., O'Donnell, M. B., Kim, H. S., Cappella, J. N., & Falk, E. B. (2017). A neural model of valuation and information virality. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1615259114

Segaert, K., Weber, K., de Lange, F. P., Petersson, K. M., & Hagoort, P. (2013). The suppression of repetition enhancement: A review of fMRI studies. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2012.11.006

Sellitto, M., Ciaramelli, E., & Di Pellegrino, G. (2010). Myopic discounting of future rewards after medial orbitofrontal damage in humans. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2516-10.2010

Shamosh, N. A., & Gray, J. R. (2008). Delay discounting and intelligence: A meta-analysis. *Intelligence*, *36*(4), 289–305.

Shohamy, D., & Daw, N. D. (2015). Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences*. https://doi.org/10.1016/j.cobeha.2015.08.010

Smith, A., Douglas Bernheim, B., Camerer, C. F., & Rangel, A. (2014). Neural activity reveals

preferences without choices. *American Economic Journal: Microeconomics*, *6*(2), 1–36. https://doi.org/10.1257/mic.6.2.1

Stillman, P. E., Lee, H., Deng, X., Rao Unnava, H., Cunningham, W. A., & Fujita, K. (2017). Neurological evidence for the role of construal level in future-directed thought. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsx022

Takahashi, H., Kato, M., Matsuura, M., Mobbs, D., Suhara, T., & Okubo, Y. (2009). When your gain is my pain and your pain is my gain: Neural correlates of envy and schadenfreude. *Science*, *323*(5916), 937–939. https://doi.org/10.1126/science.1165604

Takahashi, Y. K., Chang, C. Y., Lucantonio, F., Haney, R. Z., Berg, B. A., Yau, H. J., … Schoenbaum, G. (2013). Neural Estimates of Imagined Outcomes in the Orbitofrontal Cortex Drive Behavior and Learning. *Neuron*. https://doi.org/10.1016/j.neuron.2013.08.008

Tamir, D. I., & Mitchell, J. P. (2011). The default network distinguishes construals of proximal versus distal events. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn_a_00009

Trope, Y., & Liberman, N. (2010). Construal-Level Theory of Psychological Distance. *Psychological Review*. https://doi.org/10.1037/a0018963

Turner, B. (2010). A comparison of methods for the use of pattern classification on rapid event-related fMRI data. In *Poster session presented at the Annual Meeting of the Society for Neuroscience, San Diego, CA*.

Tusche, A., Bode, S., & Haynes, J.-D. (2010). Neural responses to unattended products predict later consumer choices. *Journal of Neuroscience*, *30*(23), 8024–8031.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Tymula, A., & Glimcher, P. (2016). *Expected Subjective Value Theory (ESVT): A Representation of Decision Under Risk and Certainty*. *SSRN*. https://doi.org/10.2139/ssrn.2783638

Venkatraman, V., Dimoka, A., Pavlou, P. A., Vo, K., Hampton, W., Bollinger, B., … Winer, R. S. (2014). *Predicting Advertising Success Beyond Traditional Measures: New Insights from Neurophysiological Methods and Market Response Modeling*. *Ssrn* (Vol. 52). American Marketing Association. https://doi.org/10.2139/ssrn.2498095

Wager, T. D., Atlas, L. Y., Leotti, L. A., & Rilling, J. K. (2011). Predicting individual differences in placebo analgesia: Contributions of brain activity during anticipation and pain experience. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.3420-10.2011

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C. W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa1204471

Weber, E. U., Shafir, S., & Blais, A. R. (2004). Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. *Psychological Review*, *111*(2), 430–445.

Wei, X. X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain “anti-Bayesian” percepts. *Nature Neuroscience*. https://doi.org/10.1038/nn.4105

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*. https://doi.org/10.2307/1912934

Yamada, H., Louie, K., Tymula, A., & Glimcher, P. W. (2018). Free choice shapes normalized value signals in medial orbitofrontal cortex. *Nature Communications*. https://doi.org/10.1038/s41467-017-02614-w

Yi, R., Stuppy-Sullivan, A., Pickover, A., & Landes, R. D. (2017). Impact of construal level manipulations on delay discounting. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0177240

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0701408104

Young, L., Dodell-Feder, D., & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2010.05.012